Stein's Method

Matthew Kwan

Supervisor: Catherine Greenhill

School of Mathematics,

The University of New South Wales

May 26, 2014

Submitted in partial fulfillment of the requirements of the degree of

Bachelor of Science with Honours

Acknowledgements

Contents

Issue 0.1. Things remaining:

- size-bias coupling
- address and remove comments
- fix overfull hboxes and large empty spaces
- (only if time) law of small numbers

1 Introduction

The *Central Limit Theorem*, roughly speaking, says that sums of independent, identically distributed random variables are "approximately normal", with the approximation improving as the number of terms in the sum increases. Similarly, the *Poisson Limit Theorem* says that the number of occurences of independent "rare" events over a given time period is "approximately Poisson". These two limit theorems are archetypal examples of a large class of related results in statistics and probabilistic combinatorics.

These types of results are generally formally stated asymptotically, with a particular type of convergence called *weak convergence*, or *convergence in distribution*. Such results are very powerful, but an obvious shortcoming is that they do not "quantify" the convergence. It may be that a sequence of random variables is "asymptotically normal", but we cannot say that any particular random variable in that sequence is individually "close to normal". Even if we are only interested in asymptotic results, it can be problematic that we cannot say convergence is "uniform" in any sense.

The solution to both these problems is to define a distance metric between probability distributions, and study convergence with respect to this metric. Fortunately, weak convergence is in fact convergence with respect to a topology, and this topology is metrizable. That is, there is a metric on the set of probability distributions such that convergence in this metric space is equivalent to weak convergence.

Stein's method was introduced by Charles Stein in 1972. It is most generally a method for approximating expected values. However, at least in probabilistic combinatorics, Stein's method has proved especially useful for bounding the distance between probability distributions, in a variety of metrics consistent with weak convergence. Such bounds can be used to quantify existing limit theorems, and can also be used as a tool to prove purely asymptotic results.

In this thesis, we will first set out the theoretical groundwork for Stein's method, including an overview of basic probability theory and a discussion of weak convergence (Chapter 2). We then present a very general framework for the Stein's method (Chapter 3) before specializing to the Normal and Poisson cases (Chapters 4 and 5). We finally outline a few particular ways

of applying the method (Chapters 6 and 7). We include a number of specific examples and applications throughout.

Unless stated otherwise, all proofs are "original" in that I came up with them, although many are quite straightforward and probably exist elsewhere. For the proofs that I have adapted from other sources (which are all cited), I tried to add value with clearer explanation and filling in skipped steps, or by adapting (and simplifying) the proof for a special case.

1.1 Notation and Assumed Knowledge

For this thesis, the set of natural numbers $\mathbb{N}$ includes zero. We write $\mathbf{1}_A$ for the characteristic function of a set A; that is, $\mathbf{1}_A(x) = 1$ if $x \in A$, otherwise $\mathbf{1}_A(x) = 0$. The function f restricted to the set A is denoted $f|_A$. The falling factorial $n(n-1)\dots(n-k+1)$ is denoted $(n)_k$. Finally, $[k]$ denotes the set $\{1, \dots, k\}$.

Unless otherwise specified, all asympotics are as $n \to \infty$. Apart from standard asymptotic notation, $f \sim g$ means $f = g(1 + o(1))$. We will occasionally use big-oh notation non-asymptotically, to avoid writing constant factors.

I will occasionally use results from analysis without proof. I will usually refer to a numbered theorem in Rudin's *Real and Complex Analysis* [Rud66] or *Functional Analysis* [Rud73] when doing so.

I will also assume some familiarity with basic graph theory definitions and terminology. The reader may refer to a textbook such as [Die00] if necessary. We also specify some conventions here. Unless stated otherwise, graphs are labelled. That is, they are distinguished even within isomorphism classes. A graph may not have loops or multiple edges; an object which is allowed to have loops and/or multiple edges will be called a multigraph. See Figure 1. We write $G \subseteq G'$ to indicate that G is a (not necessarily induced) subgraph of G'.

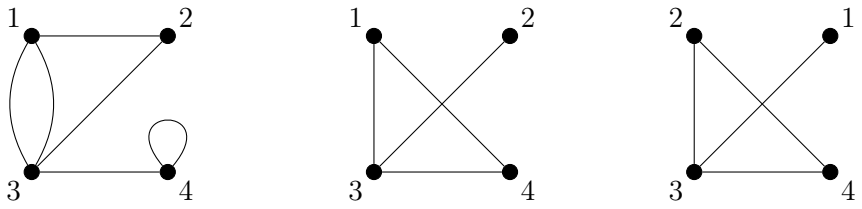

Figure 1: From left to right: a multigraph that is not a graph, and two different graphs on the vertex set $\{1, 2, 3, 4\}$.

2 General Probability Theory

For many combinatorial applications, an informal understanding of probability theory (often considering only discrete spaces) will suffice. However, in this thesis a rigorous foundation in probability theory will be useful. This section will therefore assume knowledge of basic measure theory; see, for example, [Rud66]. However, where possible, we will note any simplifications that arise from assuming discreteness.

No knowledge of probability theory is assumed; the first few subsections will briefly review the foundations of probability. The reader may nevertheless want to refer to a probability theory book such as [Kal02] for some additional detail and further reading.

2.1 Probability Spaces

Definition 2.1. A *probability space* is a measure space $(\Omega, \mathcal{A}, \mathbb{P})$ with $\mathbb{P}(\Omega) = 1$. In this case we say $\mathbb{P}$ is a *probability measure*, and denote the set of all probability measures on $(\Omega, \mathcal{A})$ by $\mathcal{P}(\Omega, \mathcal{A})$ or $\mathcal{P}(\Omega)$ if there is no ambiguity.

Remark 2.2. For our purposes Ω will often be countable, with $\mathcal{A}$ as the power set of Ω. In this case $\mathbb{P}$ is uniquely defined by the probabilities $\mathbb{P}(\omega) := \mathbb{P}(\{\omega\})$, for each $\omega \in \Omega$. We will discuss specific probability spaces on sets of combinatorial objects in Section 2.7.

Definition 2.3. An *event* in a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is a measurable set $A \in \mathcal{A}$.

For an event A, $\mathbb{P}(A)$ is interpreted as the "probability that A occurs". Events will usually be of

the form $A = \{\omega \in \Omega : P(\omega) \text{ holds}\}$, where $P(\omega)$ is some property of an object ω. For clarity, we often abuse notation slightly and write $\mathbb{P}(P \text{ holds})$ instead of $\mathbb{P}(\{\omega \in \Omega : P(\omega) \text{ holds}\})$.

2.2 Random Elements

Definition 2.4. A *random element* is a measurable function $X : (\Omega_1, \mathcal{A}_1) \to (\Omega_2, \mathcal{A}_2)$ between measurable spaces. If $\Omega_2 = \mathbb{R}^n$ for some $n \in \mathbb{N}$, with $\mathcal{A}_2$ the Borel σ-algebra on $\mathbb{R}^n$, then we say X is a *random vector*. A one-dimensional random vector is a *random variable*. If Ω_2 is countable then we say X is *discrete*.

Remark 2.5. Especially in combinatorial spaces, Ω_1 is often countable. In this case, any function from a probability space $\left(\Omega_1, 2^{\Omega_1}, \mathbb{P}\right)$ is measurable.

To interpret a random variable, we need a probability measure $\mathbb{P}$ on the underlying measurable space $(\Omega_1, \mathcal{A}_1)$ (often, this will be implicit). Then, $\mathbb{P}(X \in A) = \mathbb{P}(\{\omega \in \Omega_1 : X(\omega) \in A\})$ is the probability that X takes a value in the set A. Often, we will only be interested in such probabilities: that is, we do not care about the realization of a random variable as a function on an underlying probability space. This motivates the following definition:

Definition 2.6. Suppose X is a random element which takes values in the measurable space $(\Omega, \mathcal{A})$. The *distribution* (or *law*) $\mathcal{L}_X$ of X with respect to an underlying probability $\mathbb{P}$ is the pushforward measure with respect to X. That is, it is a probability measure defined by $\mathcal{L}_X(A) = \mathbb{P}\big(X^{-1}(A)\big)$ for $A \subseteq \mathcal{A}$. Also, we occasionally use the notation $\mathcal{L}(X) := \mathcal{L}_X$ for ease of reading.

It is worth noting that in fact any probability measure is the distribution of some random element, so we can define a probability distribution abstractly and then assert the existence of a random variable with that distribution. To see this, note that given a probability measure $\mathcal{L} \in \mathcal{P}(\Omega)$, we can choose $X = \mathrm{id}_\Omega$ to have $\mathcal{L}_X = \mathcal{L}$ with respect to the underlying probability measure $\mathcal{L}$. We also use the notation $X \in \mathcal{L}$ to indicate that X has distribution $\mathcal{L}$.

Example 2.7. The *normal distribution* with parameters μ and σ is denoted $\mathcal{N}(\mu, \sigma)$ or $\mathcal{N}_{\mu,\sigma}$ and is defined by

$$\mathcal{N}_{\mu,\sigma}(B) = \frac{1}{\sigma\sqrt{2\pi}} \int_B e^{-\frac{(x-\mu)^2}{2\sigma^2}} \,\mathrm{d}\, x$$

for any Borel set B. We can use the Gaussian integral to show that this is indeed a probability measure, with $\mathcal{N}_{\mu,\sigma}(\mathbb{R}) = 1$.

The normal distribution will be of particular importance in this thesis, so we make some remarks. Note that if $X \in \mathcal{N}(\mu, \sigma)$ then $\mathbb{E}X = \mu$ and $\operatorname{Var} X = \sigma^2$ (this second fact can be established using integration by parts and the Gaussian integral). Also, note that $\frac{X-\mu}{\sigma} \in \mathcal{N}(0,1)$. So, we will sometimes only consider the *standard normal* distribution $\mathcal{N}(0,1)$, with the understanding that it is very easy to transfer results to other normal distributions.

Example 2.8. The *Poisson distribution* with parameter λ is denoted $\mathrm{Po}(\lambda) = \mathrm{Po}_\lambda$; this is defined by

$$\mathrm{Po}_\lambda(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

for all $k \in \mathbb{N}$ (see Remark 2.2). The Taylor expansion of the exponential can be used to show that this indeed defines a probability measure.

It is in general a little tricky to define "the set of values a random element can take", but this is straightforward in the discrete case.

Definition 2.9. The *support* of a discrete random element X is the set

$$\operatorname{supp}(X) = \{k \in \Omega : \mathbb{P}(X = k) > 0\}.$$

2.3 Dependence and Coupling

Definition 2.10. Suppose X and X' are random elements $(\Omega_1, \mathcal{A}_1, \mathbb{P}) \to (\Omega_2, \mathcal{A}_2)$. We say that X and X' are *independent* if

$$\mathbb{P}(X \in A_2)\mathbb{P}\big(X' \in A_2\big) = \mathbb{P}\big(X \in A_2 \text{ and } X' \in A_2\big)$$

for all $A_2 \in \mathcal{A}_2$. If $A_1, A_1' \in \mathcal{A}$ then we analogously say A_1 and A_1' are independent if

$$\mathbb{P}(A_1)\mathbb{P}\big(A_1'\big) = \mathbb{P}\big(A_1 \cap A_1'\big).$$

We can similarly say an event is independent of a random element. If two objects are not independent, then we say they are *dependent*.

Intuitively, two objects are dependent if information about one object can give information about the other. For example, we might be interested in the probability of an event A, under the assumption that the event A' occurs.

Definition 2.11. The *conditional probability* of an event A, given an event A' which has nonzero probability, is $\mathbb{P}(A|A') = \mathbb{P}(A \cap A')/\mathbb{P}(A')$.

We have $\mathbb{P}(A|A') = \mathbb{P}(A)$ if and only if A and A' are independent.

We can also condition random elements on an event.

Definition 2.12. Suppose $X : (\Omega_1, \mathcal{A}_1, \mathbb{P}) \to (\Omega_2, \mathcal{A}_2)$ is a random element, and $A_1 \in \mathcal{A}_1$ is an event with nonzero probability. Then the *distribution of X conditioned on A_1* is denoted by $\mathcal{L}_{X|A_1}$ and defined by $\mathcal{L}_{X|A_1}(A_2) = \mathbb{P}(X \in A_2|A_1)$ for $A_2 \in \mathcal{A}_2$.

Given a finite collection of measure spaces $(\Omega_1, \mathcal{A}_1, \mu_1), \ldots, (\Omega_n, \mathcal{A}_n, \mu_n)$, recall the construction of the product measure space $(\Omega, \mathcal{A}, \mu) := (\prod_{i=1}^n \Omega_i, \bigotimes_{i=1}^n \mathcal{A}_i, \prod_{i=1}^n \mu_i)$ (see [Rud66, Chapter 8]). If a random element takes values in a product space then each component is measurable, and conversely if the components of a random tuple are measurable then that tuple is measurable in the product space. So, we can make the following definitions:

Definition 2.13. Given random elements $X_1, \ldots, X_n$ on the same underlying probability space, $\mathcal{L}(X_1, \ldots, X_n) := \mathcal{L}((X_1, \ldots, X_n))$ is called the *joint distribution* of $X_1, \ldots, X_n$. Conversely, given a random tuple $(X_1, \ldots, X_n)$, each $\mathcal{L}(X_i)$ is called a *marginal distribution*.

Suppose we have two distributions of random elements $\mathcal{L}(X_1)$ and $\mathcal{L}(X_2)$. *Coupling* is the technique of constructing a random ordered pair (X_1, X_2) which realizes the given distributions as marginal distributions. Usually this is done by specifying the joint distribution $\mathcal{L}(X_1, X_2)$.

The idea is that coupling creates a particular kind of dependence between X_1 and X_2 that allows us to compare the two distributions. Often, we are able to make conclusions about the

distributions $\mathcal{L}(X_i)$ which are independent of their specific realizations as random elements in the coupling.

Example 2.14. Let X satisfy $\mathcal{L}_X(0) = \frac{1}{2}$ and $\mathcal{L}_X(1) = \frac{1}{2}$, and let Y satisfy $\mathcal{L}_Y(0) = \frac{1}{3}$ and $\mathcal{L}_Y(1) = \frac{2}{3}$. Then we can couple X and Y into a distribution $\mathcal{L}_{X,Y}$ on $\{0,1\}^2$, defined by $\mathcal{L}_{X,Y}(0,0) = \frac{1}{3}$, $\mathcal{L}_{X,Y}(0,1) = \frac{1}{6}$, $\mathcal{L}_{X,Y}(1,0) = 0$ and $\mathcal{L}_{X,Y}(1,1) = \frac{1}{2}$. In this coupling, whenever $Y = 0$ we have $X = 0$, which allows us to make the observation $\mathcal{L}_X(0) \geq \mathcal{L}_Y(0)$. Of course, this was already obvious, but hopefully this example illustrates the principle that coupling random variables allows us to make direct comparisons between their distributions.

2.4 Expected Value

Definition 2.15. The *expected value* of a random variable X (or its distribution) is the measure-theoretic integral $\mathbb{E}X = \int x \,\mathrm{d}\,\mathcal{L}_X(x)$. It can be intuitively understood as the "average" value that random variable takes. The *variance* of X is a measure of how much variation there is from the expected value; it is defined by $\operatorname{Var} X = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$.

Remark 2.16. For a random variable X that takes integer values, our definition of expected value is equivalent to the well-known formula $\mathbb{E}X = \sum_{x\in\mathbb{Z}} x\,\mathbb{P}(X = x)$.

Remark 2.17. If X is a random variable that can be interpreted as counting the number of objects that satisfy some property, then we can express X as a sum of indicator variables $\sum_i \mathbf{1}_{A_i}$, where A_i is the event that the ith object satisfies our property. Noting that $\mathbb{E}$ is linear, we have $\mathbb{E}X = \sum_i \mathbb{E}\,\mathbf{1}_{A_i} = \sum_i \mathbb{P}A_i$. So, in order to compute the expectation of X we just need to compute the probability that each object satisfies our required property.

If we fix a particular underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$, we can also equivalently view expectation as a linear functional on the space of integrable functions (i.e. random variables): $\mathbb{E}f = \int f(\omega)\,\mathrm{d}\,\mathbb{P}$. Sometimes we will define a new probability space $(\Omega, \mathcal{A}, \mathbb{P}')$ on an existing measurable space. In this case we will write $\mathbb{E}_{\mathbb{P}'}$ to indicate expectation with respect to the measure $\mathbb{P}'$, to avoid ambiguity. We can also define the expectation functional of a random variable $\mathbb{E}_X := \mathbb{E}_{\mathcal{L}(X)}$, so that $\mathbb{E}_X f = \mathbb{E}f(X)$.

In fact, probability measures are uniquely determined by their expectation functional, because $\mathbb{E}_{\mathbb{P}} \mathbf{1}_A = \mathbb{P}(A)$ for all events A. It will be an important fact for later that much weaker classes than $\{\mathbf{1}_A : A \text{ is an event}\}$ can distinguish expectation operators.

Definition 2.18. A set of real functions $\mathcal{H}$ is a *determining class* if $\mathbb{P}_1 = \mathbb{P}_2$ whenever it is true that $\mathbb{E}_{\mathbb{P}_1} h = \mathbb{E}_{\mathbb{P}_2} h$ for all $h \in \mathcal{H}$.

Another important concept later will be the idea of conditional expectation.

Definition 2.19. The expected value of a random variable with distribution $\mathcal{L}_{X|A_1}$ is called the *conditional expected value of X given A_1* and is denoted $\mathbb{E}[X|A_1]$.

We can also define conditional expectation with respect to another random variable. If X_1 and X_2 are random variables defined on the same underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$, then the sets $X_2^{-1}(B)$ for Borel B comprise a sub-σ-algebra $\sigma(X_2)$ of $\mathcal{A}$. Then, $\mu : A' \mapsto \mathbb{E}[X_1 \mathbf{1}_{A'}]$ is a signed measure on $\sigma(X_2)$ that is absolutely continuous with respect to the restriction of $\mathbb{P}$ to $\mathcal{A}'$. By the Radon-Nikodym Theorem (see [Rud66, Theorem 6.10]) there is an $\mathcal{A}'$-measurable random variable $\mathbb{E}[X_1|X_2]$ that satisfies $\mathbb{E}[X_1 \mathbf{1}_{A'}] = \mathbb{E}[\mathbb{E}[X_1|X_2] \mathbf{1}_{A'}]$ for all A' in $\mathcal{A}'$. This random variable is almost uniquely defined: for any two choices of $\mathbb{E}[X_1|X_2]$, the probability that they differ is zero.

Definition 2.20. The random variable $\mathbb{E}[X_1|X_2]$ as defined above is called the *conditional expectation of X_1 with respect to X_2*. We can also view conditional expectation as a linear operator between functions: we define $\mathbb{E}^{X_2}$ by $X_1 \mapsto \mathbb{E}[X_1|X_2]$.

Remark 2.21. This definition generalizes the previous definition of expectation conditioned on an event: if $\omega \in A$ and $\mathbb{P}(A) > 0$ then $\mathbb{E}[X|\mathbf{1}_A](\omega) = \mathbb{E}[X|A]$.

Remark 2.22. Note that if X_2 is discrete then we do not need to invoke Radon-Nikodym. We can define $\mathbb{E}[X_1|X_2]$ by

$$\mathbb{E}[X_1|X_2](\omega) = \mathbb{E}[X_1|X_2 = X_2(\omega)]$$

for all $\omega \in \Omega$ with $\mathbb{P}(X_2 = X_2(\omega)) > 0$; this defines $\mathbb{E}[X_1|X_2]$ up to a set of probability zero.

We finally present some simple consequences of the definition of conditional expectation.

Proposition 2.23 (Tower Law of Expectation)**.** *Suppose X_1, X_2 and X_3 are random variables defined on the same underlying probability space $(\Omega, \mathcal{A})$. Further, suppose X_3 is $\sigma(X_2)$-measurable (intuitively, this means X_3 is dependent only on the information from X_2). Then*

(i) *(Tower Law)* $\mathbb{E}^{X_3}\big[\mathbb{E}^{X_2} X_1\big] = \mathbb{E}^{X_3}[X_1]$. *In particular, taking $X_3 \equiv 1$, we have* $\mathbb{E}\big[\mathbb{E}^{X_2} X_1\big] = \mathbb{E}[X_1]$.

(ii) *(Contractivity)* $\mathbb{E}\big|\mathbb{E}^{X_2} X_1\big| \leq \mathbb{E}|X_1|$.

Proof. Recall the definition of conditional expectation. For all $A \in \sigma(X_3) \subseteq \sigma(X_2)$ we have

$$\mathbb{E}[\mathbb{E}[X_1|X_3]\,\mathbf{1}_A] = \mathbb{E}[X_1\,\mathbf{1}_A] = \mathbb{E}[\mathbb{E}[X_1|X_2]\,\mathbf{1}_A].$$

Since $\mathbb{E}[X_1|X_3]$ is $\sigma(X_3)$-measurable, the tower law is proved. Next, let $A = \big\{\mathbb{E}^{X_2} X_1 \geq 0\big\}$, so

$$\mathbb{E}\big|\mathbb{E}^{X_2} X_1\big| = \mathbb{E}\big[\mathbf{1}_A\,\mathbb{E}^{X_2} X_1\big] - \mathbb{E}\big[\mathbf{1}_{\Omega\setminus A}\,\mathbb{E}^{X_2} X_1\big] = \mathbb{E}[\mathbf{1}_A\,X_1] - \mathbb{E}\big[\mathbf{1}_{\Omega\setminus A}\,X_1\big] \leq \mathbb{E}|X_1|,$$

proving contractivity. □

2.5 Markov Chains

We discuss Markov Chains only superficially, to the extent that is required for Chapter 7. The theory of Markov Chains (and Markov Processes in general) is a very rich area of probability theory, see for example [Kal02, Chapter 7].

For our purposes, a (time-homogeneous, two-sided) *Markov Chain* is a sequence of discrete random elements $(X_i)_{i\in\mathbb{Z}}$ taking values on a common set Ω (the *state space*), such that

$$\mathbb{P}(X_{i+1} = y|X_i = x) = \mathbb{P}(X_i = y|X_{i-1} = x) \tag{2.1}$$

for all $x, y \in \Omega$, $n \in \mathbb{Z}$. The probability in (2.1) is called the *transition probability* from x to y, and is denoted p_{xy} or $p(x, y)$. We can think of $(X_i)_{i\in\mathbb{Z}}$ as a process evolving in time in such a way that, given the current state, the next state does not depend on past states or the current point in time.

Definition 2.24. A Markov Chain $(X_i)_{i\in\mathbb{Z}}$ is stationary with respect to the distribution $\mathcal{L}(X)$ if $\mathcal{L}(X_i) = \mathcal{L}(X)$ for all $i \in \mathbb{Z}$.

Define $\pi_x = \mathbb{P}(X = x)$. Note that if $(X_i)_{i\in\mathbb{Z}}$ is stationary with respect to $\mathcal{L}(X)$, then we must have

$$\pi_y = \mathbb{P}(X_i = y) = \sum_{x\in\Omega} \mathbb{P}(X_i = y,\, X_{i-1} = x) = \sum_{x\in\Omega} \pi_x p_{xy} \tag{2.2}$$

for all $y \in \Omega$. In fact, the distribution $\mathcal{L}(X_i)_{i\in\mathbb{Z}}$ of a stationary Markov Chain is uniquely defined by a stationary distribution $\mathcal{L}_X$ and a set of transition probabilities $\{p_{xy}\}_{x,y\in\Omega}$ which satisfy (2.2) for all $y \in \Omega$ and $\sum_{y\in\Omega} p_{xy} = 1$ for all $x \in \Omega$.

Definition 2.25. A stationary Markov Chain $(X_i)_{i\in\mathbb{Z}}$ is *reversible* if $\mathcal{L}(X_i)_{i\in\mathbb{Z}} = \mathcal{L}(X_{-i})_{i\in\mathbb{Z}}$.

Reversibility means that a Markov Chain can be equally well understood as evolving forward or backward in time. As before, let $\pi_x = \mathbb{P}(X = x)$ for the stationary distribution $\mathcal{L}_X$. Then, reversibility is equivalent to the condition $\pi_x p_{xy} = \pi_y p_{yx}$ for all $x, y \in \Omega$.

Definition 2.26. A Markov Chain $(X_i)_{i\in\mathbb{Z}}$ is *irreducible* if for all $x, y \in \Omega$ there is $t \in \mathbb{Z}$ such that $\mathbb{P}(X_{i+t} = y | X_i = x) > 0$.

Irreducibility means that every state can be accessed from every other state. Irreducibility is equivalent to the condition that for any $x, y \in \Omega$, there is a finite sequence $x = x_0, x_1, \ldots, x_k = y$ with $p(x_i, x_{i+1}) > 0$ for all $i < k$.

Proposition 2.27. *Let G be a finite connected edge-weighted multigraph. Suppose that each edge weight is positive, and the sum of the weights at each vertex is Δ. Then G defines a reversible Markov Chain on its vertex set V, which is stationary with respect to the uniform distribution on V.*

Proof. For $x, y \in V$, let $p_{xy} = p_{yx}$ be the sum of the weights of all edges between x and y, divided by Δ. By construction, $\sum_{y\in\Omega} p_{xy} = 1$. Let $\mathcal{L}_X$ be the uniform distribution on V, and let $\pi_x = \mathcal{L}_X(x)$. Then,

$$\pi_y = \frac{1}{|V|} = \frac{1}{|V|} \sum_{x\in V} p_{xy} = \sum_{x\in V} \pi_x p_{xy}$$

for all $y \in V$, and also

$$\pi_x p_{xy} = \frac{1}{|V|} p_{xy} = \pi_y p_{yx}$$

for each $x, y \in V$. We conclude that the transition probabilities $\{p_{xy}\}_{x,y \in V}$ define a reversible Markov Chain stationary with respect to $\mathcal{L}_X$. □

Example 2.28. Let G be a cycle of length n, with a weight of 1 on each edge. The corresponding Markov Chain is a random walk: at each time-step, there is equal probability of moving one step clockwise or counterclockwise around the cycle.

2.6 The Weak Topology on Probability Measures

Definition 2.29. Let $(X_n)_{n \in \mathbb{N}}$ be a sequence of random variables, not necessarily on the same underlying probability space. We say X_n *converges in distribution* to a random variable X if $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for all bounded continuous functions $f : \mathbb{R} \to \mathbb{R}$. Alternatively, we say $\mathcal{L}(X_n)$ converges *weakly* to $\mathcal{L}(X)$, or simply $\mathcal{L}(X_n) \to \mathcal{L}(X)$.

This is only one of a number of equivalent natural definitions for convergence in distribution.

Definition 2.30. The *distribution function* F_X of a random variable X is defined by $F_X(x) = \mathbb{P}(X \leq x)$ for all $x \in \mathbb{R}$.

Theorem 2.31. *For a sequence of random variables $(X_n)_{n \in \mathbb{N}}$ and a random variable X, the following are equivalent.*

(i) $\mathcal{L}(X_n) \to \mathcal{L}(X)$*;*

(ii) $\liminf_{n \to \infty} \mathcal{L}_{X_n}(U) \geq \mathcal{L}_X(U)$ *for every open $U \subseteq \mathbb{R}$;*

(iii) $F_{X_n}(x) \to F_X(x)$ *for all x such that F_X is continuous at x;*

(iv) $\mathbb{E}e^{itX_n} \to \mathbb{E}e^{itX}$ *for all $t \in \mathbb{R}$.*

The equivalence of Conditions (i) to (iii) is (part of) a well-known and relatively elementary result called the Portmanteau Theorem ([Kal02, Theorem 3.25]). The equivalence of Conditions (i) and (iv) is called *Lévy's Continuity Theorem* ([Kal02, Theorem 4.3]).

If X and each X_n are integer random variables, then Condition (iii) reduces to the condition that $\mathbb{P}(X_n = k) \to \mathbb{P}(X = k)$ for all k. This characterization is usually used to prove the Poisson Limit Theorem.

Classically, distributional convergence results are usually proved with Lévy's continuity theorem and Fourier analysis. For example, this approach is used to prove the Central Limit Theorem. In combinatorial applications, convergence in distribution is often proved with the "method of moments" (which is actually the Fourier analytic method in disguise): under certain conditions, $\mathcal{L}(X_n) \to \mathcal{L}(X)$ if $\mathbb{E}X_n^k \to \mathbb{E}X^k$ for all k. Convergence in distribution can also sometimes be inferred from stronger forms of convergence when X and all the X_n are coupled to the same underlying space.

In functional analysis terms, note that expectation operators are bounded linear functionals on the space of real bounded continuous functions. Then, $\mathcal{L}(X_n) \to \mathcal{L}(X)$ just means that $\mathbb{E}_{X_n} \to \mathbb{E}_X$ in the weak-star topology. The subspace corresponding to the set of probability distributions $\mathcal{P}(\mathbb{R})$ is confusingly called the *weak topology* on probability distributions.

Although $C_b(\mathbb{R})^*$ is not metrizable, the unit ball (in operator norm) of $C_b(\mathbb{R})^*$ is in fact metrizable (see [Rud73, Theorems 3.15 and 3.16]). Every expectation functional $\mathbb{E}$ has unit operator norm because $\mathbb{E}|h| \leq \mathbb{E}|1| = 1$ for h with unit uniform norm. So, the weak topology is metrizable.

Definition 2.32. Let $\mathcal{H}$ be a determining class of real measurable "test" functions that are uniformly absolutely bounded. Define $d_{\mathcal{H}} : \mathcal{P}(\mathbb{R})^2 \to \mathbb{R}^+$ by

$$d_{\mathcal{H}}(\mathbb{P}_1, \mathbb{P}_2) = \sup_{h \in \mathcal{H}} |\mathbb{E}_{\mathbb{P}_1} h - \mathbb{E}_{\mathbb{P}_2} h|.$$

For random variables X_1, X_2, we write $d_{\mathcal{H}}(X_1, X_2)$ instead of $d_{\mathcal{H}}(\mathcal{L}(X_1), \mathcal{L}(X_2))$.

Proposition 2.33. *Each $d_{\mathcal{H}}$ as defined above is a metric.*

Proof. Since the functions in $\mathcal{H}$ are bounded, $\mathbb{E}_{\mathbb{P}_1} h$ is well-defined for every $h \in \mathcal{H}$. Since the bound is uniform, the supremum in the definition of $d_{\mathcal{H}}$ is finite. It is immediate that $d_{\mathcal{H}}$ is

non-negative, symmetric and satisfies the triangle inequality. Finally, $d_{\mathcal{H}}(\mathbb{P}_1, \mathbb{P}_2) = 0$ implies that $\mathbb{E}_{\mathbb{P}_1} h = \mathbb{E}_{\mathbb{P}_2} h$ for all $h \in \mathcal{H}$. Since $\mathcal{H}$ is a determining class, $\mathbb{P}_1 = \mathbb{P}_2$. □

Definition 2.34. We define some special cases of $d_{\mathcal{H}}$.

- If $\mathcal{H}_{\mathrm{K}} = \left\{ \mathbf{1}_{(-\infty, x]} : x \in \mathbb{R} \right\}$ then $d_{\mathrm{K}} := d_{\mathcal{H}_{\mathrm{K}}}$ is called the *Kolmogorov metric.*
- Let $\mathcal{H}_{\mathrm{BL}}$ be the set of functions h that are absolutely bounded by 1 (that is, $|h(x)| \leq 1$ for all $x \in \mathbb{R}$), and have Lipschitz constant at most 1 (that is, $|h(x_1) - h(x_2)| \leq |x_1 - x_2|$ for all $x_1, x_2 \in \mathbb{R}$). Then, $d_{\mathrm{BL}} := d_{\mathcal{H}_{\mathrm{BL}}}$ is called the *Bounded Lipschitz metric.*
- If $\mathcal{H}_{\mathrm{TV}}$ is the set of functions $\mathbf{1}_B$ for Borel B, then $d_{\mathrm{TV}} := d_{\mathcal{H}_{\mathrm{TV}}}$ is called the *total variation metric.*
- If $\mathcal{H}_{\mathrm{W}}$ is the set of functions with Lipschitz constant at most 1, then $d_{\mathrm{W}} := d_{\mathcal{H}_{\mathrm{W}}}$ is called the *Wasserstein metric.* However, since $\mathcal{H}$ is not uniformly bounded, d_{W} is not strictly speaking a metric on $\mathcal{P}(\mathbb{R})$; the Wasserstein metric can only be used to compare distributions with finite first moment, as we will see below.

Proposition 2.35. *The "metrics" in Definition 2.34 are actually metrics.*

Proof. First, note that if X_1 and X_2 have finite first moment then for all $h \in \mathcal{H}_{\mathrm{W}}$,

$$|\mathbb{E}h(X_1) - \mathbb{E}h(X_2)| \leq |(\mathbb{E}|X_1| + h(0)) - (\mathbb{E}|X_2| + h(0))| \leq \mathbb{E}|X_1| + \mathbb{E}|X_2|.$$

So,

$$d_{\mathrm{W}}(X_1, X_2) < \infty.$$

Recalling that every probability measure is the distribution of some random variable, we have proved that $d_{\mathrm{W}} : \mathcal{P}(\mathbb{R})^2 \to \mathbb{R}^+$ is well-defined.

Now, by Proposition 2.33 we just need to check that each of $\mathcal{H}_{\mathrm{K}}$, $\mathcal{H}_{\mathrm{BL}}$, $\mathcal{H}_{\mathrm{TV}}$, $\mathcal{H}_{\mathrm{W}}$ are determining classes. Let $\mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P}(\mathbb{R})$ satisfy $\mathbb{E}_{\mathbb{P}_1} h = \mathbb{E}_{\mathbb{P}_2} h$ for each h in $\mathcal{H}$.

If $\mathcal{H}$ is $\mathcal{H}_{\mathrm{K}}$ or $\mathcal{H}_{\mathrm{TV}}$ then $\mathbf{1}_{(-\infty, x]} \in \mathcal{H}$ for all $x \in \mathbb{R}$ so

$$\mathbb{P}_1((-\infty, x]) = \mathbb{E}_{\mathbb{P}_1} \mathbf{1}_{(-\infty, x]} = \mathbb{E}_{\mathbb{P}_2} \mathbf{1}_{(-\infty, x]} = \mathbb{P}_2((-\infty, x]).$$

Since $\{(-\infty, x] : x \in \mathbb{R}\}$ generates the Borel σ-algebra, $\mathbb{P}_1 = \mathbb{P}_2$. We have shown that $\mathcal{H}_{\mathrm{K}}$ and $\mathcal{H}_{\mathrm{TV}}$ are determining classes.

For $x \in \mathbb{R}$ and $0 < \varepsilon \leq 1$, let $h_{x,\varepsilon}$ be the continuous function which takes the value 1 on the set $(-\infty, x]$, takes the value 0 on the set $[x + \varepsilon, \infty)$, and is linearly interpolated in the range $[x, x + \varepsilon]$. Suppose $\mathcal{H}$ is $\mathcal{H}_{\mathrm{BL}}$ or $\mathcal{H}_{\mathrm{W}}$, so that each $\varepsilon h_{x,\varepsilon} \in \mathcal{H}$. It follows that $\mathbb{E}_{\mathbb{P}_1} h_{x,\varepsilon} = \mathbb{E}_{\mathbb{P}_2} h_{x,\varepsilon}$ for each $x \in \mathbb{R}$ and $0 < \varepsilon \leq 1$. For each $x \in \mathbb{R}$, note that $h_{x,1/n} \to \mathbf{1}_{(-\infty,x]}$ pointwise, and each $h_{x,1/n} \leq 1$. By the Dominated Convergence Theorem (see [Rud66, Theorem 1.34]),

$$\mathbb{P}_1((-\infty, x]) = \mathbb{E}_{\mathbb{P}_1} \mathbf{1}_{(-\infty,x]} = \lim_{n\to\infty} \mathbb{E}_{\mathbb{P}_1} h_{x,1/n} = \lim_{n\to\infty} \mathbb{E}_{\mathbb{P}_2} h_{x,1/n} = \mathbb{E}_{\mathbb{P}_2} \mathbf{1}_{(-\infty,x]} = \mathbb{P}_2((-\infty, x]),$$

so $\mathcal{H}_{\mathrm{BL}}$ and $\mathcal{H}_{\mathrm{W}}$ are determining classes, as above. □

Proposition 2.36. *The metrics in Definition 2.34 are each stronger than the weak topology.*

Proof. We show that $d_{\mathcal{H}}(X_n, X) \to 0$ implies $\mathcal{L}(X_n) \to \mathcal{L}(X)$ for each $\mathcal{H}$.

If $d_{\mathrm{K}}(X_n, X) \to 0$ or $d_{\mathrm{TV}}(X_n, X) \to 0$ then $F_{X_n} \to F_X$ uniformly, so certainly Condition (iii) of Theorem 2.31 holds.

Now, suppose $d_{\mathrm{BL}}(X_n, X) \to 0$ (this will automatically be true if $d_{\mathrm{W}}(X_n, X) \to 0$). Let $d_n = \sqrt{d_{\mathrm{BL}}(X_n, X)}$ and recall the definition of $h_{x,\varepsilon}$ from the proof of Proposition 2.35. Since $d_n h_{x,d_n} \in \mathcal{H}_{\mathrm{W}}$ for each $n \in \mathbb{N}$, we have

$$\mathbb{E}_{X_n} h_{x,d_n} - \mathbb{E}_X h_{x,d_n} \leq d_{\mathrm{W}}(X_n, X)/d_n = d_n \to 0$$

uniformly for $x \in \mathbb{R}$. Now, note that

$$F_X(x - \varepsilon) \leq \mathbb{E}_X h_{x-\varepsilon,\varepsilon} \leq F_X(x) \leq \mathbb{E}_X h_{x,\varepsilon} \leq F_X(x + \varepsilon)$$

for any random variable X. If F_X is continuous at x then

$$\begin{aligned} F_{X_n}(x) - F_X(x) &\leq (\mathbb{E}_{X_n} h_{x,d_n} - \mathbb{E}_X h_{x,d_n}) + (F_X(x + d_n) - F_X(x)) \to 0, \\ F_{X_n}(x) - F_X(x) &\geq (\mathbb{E}_{X_n} h_{x-d_n,d_n} - \mathbb{E}_X h_{x-d_n,d_n}) + (F_X(x - d_n) - F_X(x)) \to 0, \end{aligned}$$

so Condition (iii) of Theorem 2.31 holds. □

Theorem 2.37. *The bounded Lipschitz metric metrizes the weak topology.*

We will need a small lemma to prove Theorem 2.37.

Lemma 2.38. *Let $S \subseteq \mathcal{P}(\mathbb{R})$ be compact in the weak topology. For each $\varepsilon > 0$, there is $k \in \mathbb{N}$ such that $\sup_{\mathbb{P} \in S} \mathbb{P}((-k,k)^c) \leq \varepsilon$.*

Proof. Fix $\varepsilon > 0$. Suppose for the purpose of contradiction that for all $k \in \mathbb{N}$ there is $\mathbb{P}_k \in S$ with $\mathbb{P}_k((-k,k)^c) > \varepsilon$. Since S is compact, there is a subsequence $(\mathbb{P}_{k_n})_{n\in\mathbb{N}}$ of $(\mathbb{P}_k)_{k\in\mathbb{N}}$, and a measure $\mathbb{P} \in S$, with $\mathbb{P}_{k_n} \to \mathbb{P}$ as $n \to \infty$. For each $k \in \mathbb{N}$ we have $k < k_n$ for sufficiently large n, so by Condition (ii) of Theorem 2.31, we have

$$\mathbb{P}((-k,k)) \leq \liminf_{n\to\infty} \mathbb{P}_{k_n}((-k,k)) \leq \liminf_{n\to\infty} \mathbb{P}_{k_n}((-k_n,k_n)) \leq 1-\varepsilon.$$

This is a contradiction because $\mathbb{P}((-k,k)) = \mathbb{E}_{\mathbb{P}} \mathbf{1}_{(-k,k)} \to \mathbb{E}_{\mathbb{P}} 1 = 1$ as $k \to \infty$, by the Dominated Convergence Theorem (see [Rud66, Theorem 1.34]). □

Proof of Theorem 2.37. This proof is simplified from a more general proof in [Vil03, Theorem 7.12] (this simplification is in fact presented as an exercise).

Let $\mathbb{P}_n \to \mathbb{P}_\infty$ weakly, and suppose for the purpose of contradiction that $d_{\mathrm{BL}}(\mathbb{P}_n, \mathbb{P}_\infty) \not\to 0$. Then there is $\varepsilon > 0$ and a subsequence $(\mathbb{P}_{k_n})_{n\in\mathbb{N}}$ with $d_{\mathrm{BL}}(\mathbb{P}_{k_n}, \mathbb{P}_\infty) > 2\varepsilon$ for all n. To simplify notation, redefine $\mathbb{P}_n$ to be $\mathbb{P}_{k_n}$ for each n (we still have $\mathbb{P}_n \to \mathbb{P}_\infty$ weakly).

Now, by the definition of d_{BL}, for each $n \in \mathbb{N}$ there is $h_n = h_n^{(0)} \in \mathcal{H}_{\mathrm{BL}}$ with

$$|\mathbb{E}_{\mathbb{P}_n} h_n - \mathbb{E}_{\mathbb{P}} h_n| \geq d_{\mathrm{BL}}(\mathbb{P}_n, \mathbb{P}) - \varepsilon > \varepsilon.$$

For each $k \in \mathbb{Z}^+$, let $\mathcal{H}_{\mathrm{BL}}^{(k)} = \left\{h|_{[-k,k]} : h \in \mathcal{H}_{\mathrm{BL}}\right\}$. By the Arzela-Ascoli Theorem (see [Rud66, Theorem 11.28]), each $\mathcal{H}_{\mathrm{BL}}^{(k)}$ is a compact subset of $C_b([-k,k])$ with the uniform norm. So, for each $k \in \mathbb{Z}^+$, we can inductively choose a subsequence $\left(h_n^{(k)}\right)_{n\in\mathbb{N}}$ of $\left(h_n^{(k-1)}\right)_{n\in\mathbb{N}}$ such that $h_n^{(k)}|_{[-k,k]}$ converges uniformly to some $h^{(k)} \in \mathcal{H}_{\mathrm{BL}}^{(k)}$.

Note that $h_n^{(n)}|_{[-k,k]} \to h^{(k)}$ uniformly for each $k \in \mathbb{Z}^+$, so that $h_n^{(n)}$ converges pointwise to some function $h : \mathbb{R} \to \mathbb{R}$ that satisfies $h|_{[-k,k]} = h^{(k)}$ for all $k \in \mathbb{Z}^+$. Since h is bounded by 1 and has Lipschitz constant less than 1 on each $[-k,k]$, it follows that $h \in \mathcal{H}_{\mathrm{BL}}$ and in particular h is bounded and continuous. Finally, note that

$$\begin{aligned}|\mathbb{E}_{\mathbb{P}_n} h_n - \mathbb{E}_{\mathbb{P}_\infty} h_n| \leq & \left|\mathbb{E}_{\mathbb{P}_n}\left[(h_n - h)\,\mathbf{1}_{[-k,k]}\right]\right| + \left|\mathbb{E}_{\mathbb{P}_\infty}\left[(h_n - h)\,\mathbf{1}_{[-k,k]}\right]\right| \\ & + \left|\mathbb{E}_{\mathbb{P}_n}\left[(h_n - h)\,\mathbf{1}_{[-k,k]^c}\right]\right| + \left|\mathbb{E}_{\mathbb{P}_\infty}\left[(h_n - h)\,\mathbf{1}_{[-k,k]^c}\right]\right| \\ & + \left|\mathbb{E}_{\mathbb{P}_n} h - \mathbb{E}_{\mathbb{P}_\infty} h\right|\end{aligned}$$

By the definition of weak convergence, there is $N \in \mathbb{N}$ such that

$$\left|\mathbb{E}_{\mathbb{P}_n} h - \mathbb{E}_{\mathbb{P}_\infty} h\right| \leq \frac{\varepsilon}{5}$$

for all $n > N$. Since the weak topology is metrizable, $\{\mathbb{P}_n\}_{n \in \mathbb{N} \cup \{\infty\}}$ is compact so by Lemma 2.38, there is $k \in \mathbb{N}$ such that

$$\mathbb{E}_{\mathbb{P}_n}\left[(h_n - h)\,\mathbf{1}_{[-k,k]^c}\right] \leq 2\mathbb{P}_n[(-k,k)^c] \leq \frac{\varepsilon}{5}$$

for $n \in \mathbb{N} \cup \{\infty\}$. Since $\left(h_n^{(n)}\right)_{n \in \mathbb{N}}$ is a subsequence of $(h_n)_{n \in \mathbb{N}}$ and $h_n^{(n)}\,\mathbf{1}_{[-k,k]}$ converges uniformly to $h\,\mathbf{1}_{[-k,k]}$, there is $n > N$ such that

$$\left\|(h_n - h)\,\mathbf{1}_{[-k,k]}\right\|_\infty \leq \frac{\varepsilon}{5}.$$

For this n we have $|\mathbb{E}_{\mathbb{P}_n} h_n - \mathbb{E}_{\mathbb{P}_\infty} h_n| \leq \varepsilon$. This is a contradiction. □

Proposition 2.36 tells us that our selection of "special" metrics are all "consistent" with weak convergence in some way, and Theorem 2.37 tells us that convergence in the Bounded Lipschitz metric is exactly the same as weak convergence. Typically, it will be natural to work with the total variation metric for Poisson approximation, and to work with the Wasserstein metric for Normal approximation. In applications, we may be most interested in the Kolmogorov metric. Therefore, it is sometimes useful to transfer results between metrics (though, this usually results in worse constants than working directly in the desired metric).

Comment 2.1. It may be worthwhile to further discuss the Wasserstein, Kolmogorov and Total Variation topologies. In particular, Wasserstein convergence is just weak convergence plus convergence of the first moment.

Definition 2.39. If (for all $x \in \mathbb{R}$), $F_X(x) = \int_{-\infty}^{x} f_X(x)\,\mathrm{d}\,x$ for some f_X, then f_X is called the *Lebesgue density* of X, and X is called a *continuous* random variable.

If X is a continuous random variable, then $\mathbb{E}_X h = \int_{\mathbb{R}} h(x) f_X(x)\,\mathrm{d}\,x$ for measurable h. (This is immediate from the measure-theoretic definition of the integral).

Proposition 2.40. *Let X_1, X_2 be random variables. Then*

(i) $d_{\mathrm{K}}(X_1, X_2) \leq d_{\mathrm{TV}}(X_1, X_2)$*;*

(ii) $d_{\mathrm{BL}}(X_1, X_2) \leq d_{\mathrm{W}}(X_1, X_2)$*;*

(iii) *If there is $C > 0$ such that $|f_{X_2}(x)| \leq C$ for all x, then $d_{\mathrm{K}}(X_1, X_2) \leq \sqrt{2C d_{\mathrm{BL}}(X_1, X_2)}$.*

Proof. (Adapted from [Ros11, Proposition 1.2]). Items (i) and (ii) are immediate from the definitions. Then, as in the proof of Proposition 2.36,

$$\begin{aligned} F_{X_n}(x) - F_X(x) &\leq (\mathbb{E}_{X_n} h_{x,\varepsilon} - \mathbb{E}_X h_{x,\varepsilon}) + (\mathbb{E}_X h_{x,\varepsilon} - F_X(x)) \\ &\leq d_{\mathrm{BL}}(X_1, X_2)/\varepsilon + \int_x^{x+\varepsilon} h_{x,\varepsilon}\, f_X(x)\,\mathrm{d}\,x \\ &\leq d_{\mathrm{BL}}(X_1, X_2)/\varepsilon + C\varepsilon/2 \end{aligned}$$

and similarly

$$F_{X_n}(x) - F_X(x) \geq -d_{\mathrm{BL}}(X_1, X_2)/\varepsilon - C\varepsilon/2.$$

So, we can take $\varepsilon = \sqrt{2 d_{\mathrm{BL}}(X_1, X_2)/C}$ to prove Item (iii). □

Example 2.41. If $\mathcal{L}(X_2) = \mathcal{N}(0,1)$ then $f_{X_2}(x) = (2\pi)^{-1/2} e^{-x^2/2}$ so we can take $C = (2\pi)^{-1/2}$ to obtain $d_{\mathrm{K}} \leq (2/\pi)^{1/4} \sqrt{d_{\mathrm{BL}}(X_1, X_2)}$.

2.7 Random Combinatorial Structures

Definition 2.42. Given a finite space of combinatorial objects Ω, a probability space $\left(\Omega, 2^{\Omega}, \mathbb{P}\right)$ is often called a *model* of Ω.

Definition 2.43. In a probability space $\left(\Omega, 2^{\Omega}, \mathbb{P}\right)$ where Ω is finite, if $\mathbb{P}(\omega) = 1/|\Omega|$ for each $\omega \in \Omega$, then we say the space is *uniform*.

Uniform models are the simplest examples of random structures. For example, the uniform space $\mathcal{S}_n$ of permutations on n elements has $\mathbb{P}(\sigma) = 1/n!$ for each $\sigma \in S_n$. The uniform random regular graph model $\mathcal{G}_{n,d}$ is uniform on the set of all d-regular graphs on the vertex set $[n]$, though an explicit formula for the number of such graphs is not known. (Recall that a d-regular graph is a graph where each vertex has exactly d incident edges).

As an important example of a (generally) non-uniform model, the (Erdős-Rényi) binomial random graph model $\mathcal{G}_{n,p}$ has

$$\mathbb{P}(G) = p^{|E(G)|}(1-p)^{\binom{n}{2}-|E(G)|}$$

for each graph G on the vertex set $[n]$. When $p = 1/2$, we obtain the uniform model on all graphs on the vertex set $[n]$. It is unfortunate that the common notations for the random regular model and the binomial model clash, but hopefully in this thesis the intended meaning will always be very clear from context.

One way to conceptualize the binomial model is to consider a sequence of independent coin tosses, where the coin is biased to land heads with probability p. Each coin toss corresponds to a particular potential edge, and determines whether that edge is present in the final random graph. When we define more complicated random models, we will often use this kind of informal description rather than giving an explicit formula for each $\mathbb{P}(\omega)$.

3 Stein's Method in Generality

There are a number of different presentations of Stein's method, with slightly different assumptions and different (though mostly equivalent) definitions. I have decided to describe Stein's original approach [Ste86] because, in particular, it does a better job motivating the method of exchangeable pairs in Chapter 7.

Suppose we have a potentially complicated random variable X, and we believe the distribution of X is close to a "special" distribution $\mathcal{L}_0$. Then, Stein's method allows us to compare the operators $\mathbb{E}_X$ and $\mathbb{E}_0 := \mathbb{E}_{\mathcal{L}_0}$. This is sometimes directly useful for approximating statistics of X (for example, $\mathbb{P}(X \in A) = \mathbb{E}_X \mathbf{1}_A$). However, particularly for combinatorial applications, Stein's method is most often used to bound the distance $d_{\mathcal{H}}(\mathcal{L}_X, \mathcal{L}_0)$ for some $\mathcal{H}$, where the metric $d_{\mathcal{H}}$ from Definition 2.32 is defined in terms of $\mathbb{E}_X$ and $\mathbb{E}_0$.

Stein's method is motivated by the idea of a characterizing operator.

Definition 3.1. Let $\mathcal{F}_0$ be a vector space and $\mathcal{V}_0$ be a vector space of measurable functions which contains the constant functions. We say a linear operator $T_0 : \mathcal{F}_0 \to \mathcal{V}_0$ is a *characterizing operator* for the distribution $\mathcal{L}_0$ if $\operatorname{im} T_0 = \mathcal{V}_0 \cap \ker \mathbb{E}_0$. For convenience, where there is no ambiguity we will often implicitly restrict $\mathbb{E}_0$ to $\mathcal{V}_0$, so we can write $\operatorname{im} T_0 = \ker \mathbb{E}_0$.

The following proposition shows why T_0 is called a characterizing operator.

Proposition 3.2. *Suppose $T_0 : \mathcal{F}_0 \to \mathcal{V}_0$ is a characterizing operator and $\mathcal{V}_0$ is a determining class. If some random variable X satisfies $\operatorname{im} T_0 \subseteq \ker \mathbb{E}_X$, then $X \in \mathcal{L}_0$.*

Proof. If $h \in \mathcal{V}_0$, then $h - \mathbb{E}_0 h \in \ker \mathbb{E}_0 = \operatorname{im} T_0$ so $\mathbb{E}_X[h - \mathbb{E}_0 h] = 0$. That is, $\mathbb{E}_X h = \mathbb{E}_0 h$ for all $h \in \mathcal{V}_0$, which means $\mathcal{L}_X = \mathcal{L}_0$ by the definition of a determining class. □

Proposition 3.2 should already give some clue as to how Stein's method works. If $\mathbb{E}_X T_0 f = 0$ for all f means $\mathcal{L}_X = \mathcal{L}_0$, then hopefully it is true that $\mathbb{E}_X T_0 f \approx 0$ for all f means $\mathcal{L}_X \approx \mathcal{L}_0$. In practice it is a little unwieldy to show an operator is characterizing using Definition 3.1, so we'll prove an equivalent definition.

Proposition 3.3. *$T_0 : \mathcal{F}_0 \to \mathcal{V}_0$ is characterizing if and only if there is a linear operator $U_0 : \mathcal{V}_0 \to \mathcal{F}_0$ (called a* Stein transform*) such that the following two equations hold:*

$$\mathbb{E}_0 T_0 = 0_{\mathcal{F}_0}, \tag{3.1}$$

$$T_0 U_0 + \mathbb{E}_0 = \mathrm{id}_{\mathcal{V}_0}. \tag{3.2}$$

Proof. Suppose T_0 is a characterizing operator. Equation (3.1) is immediate. Let $\{h_i\}_{i\in\mathcal{I}}$ be a (Hamel) basis of $\mathcal{V}_0$. For each $i \in \mathcal{I}$ we have $h_i - \mathbb{E}_0 h_i \in \ker \mathbb{E}_0$ so there is some f_i (not necessarily unique) that solves $T_0 f_i = h_i - \mathbb{E}_0 h_i$. The operator U_0 can then be defined by $\sum_{i\in\mathcal{I}} a_i h_i \mapsto \sum_{i\in\mathcal{I}} a_i f_i$, satisfying (3.2).

Conversely, suppose (3.1) holds and U_0 exists satisfying (3.2). For $h \in \ker \mathbb{E}_0$ we have $T_0(U_0 h) = h$ and hence $h \in \operatorname{im} T_0$, so $\ker \mathbb{E}_0 \subseteq \operatorname{im} T_0$. Equation (3.1) immediately says that $\operatorname{im} T_0 \subseteq \ker \mathbb{E}_0$, so T_0 is a characterizing operator. □

Remark 3.4. In full generality (and in accordance with [Ste86]), we do not require any topological structure on $\mathcal{F}_0$ and $\mathcal{V}_0$. The proof of Proposition 3.3 uses the Axiom of Choice for the existence of a basis, and does not ensure that the Stein transform U_0 is particularly well-behaved. In practice, we will usually require that U_0 is well-behaved when we apply Stein's method (in fact, we will usually have an explicit formula for U_0).

Comment 3.1. I managed to prove a necessary and sufficient condition for U_0 to be bounded if T_0 is a Banach space operator (namely, $\ker T_0$ is complementable). I think it's probably a little off-topic to include this though.

Now, note that for any random variable X and any $h \in \mathcal{V}_0$, Equation (3.2) allows us to to make the transformation

$$\mathbb{E}_X h = \mathbb{E}_0 h + \mathbb{E}_X T_0 U_0 h. \tag{3.3}$$

The original purpose of Stein's method was to estimate some particular $\mathbb{E}_X h$. If $\mathcal{L}_0$ is a well-understood distribution then the term $\mathbb{E}_0 h$ should be easy to compute or estimate, and if the

distribution of X was "close" to $\mathcal{L}_0$, then it should be possible to show that the remainder $\mathbb{E}_X T_0 U_0 h$ is small.

For our purposes, the main use of (3.3) is to bound $d_{\mathcal{H}}(X, \mathcal{L}_0)$ for some $\mathcal{H} \subseteq \mathcal{V}_0$. For any $\mathcal{Y} \supseteq U_0\mathcal{H}$, we have

$$d_{\mathcal{H}}(X, \mathcal{L}_0) = \sup_{h \in \mathcal{H}} |\mathbb{E}_X T_0 U_0 h| \leq \sup_{f \in \mathcal{Y}} |\mathbb{E}_X T_0 f|. \tag{3.4}$$

We have reduced the problem of bounding $d_{\mathcal{H}}(X, \mathcal{L}_0)$ to that of bounding $|\mathbb{E}_X T_0 f|$ (uniformly over $f \in \mathcal{Y}$). Especially in the cases where $\mathcal{L}_0$ is normal or Poisson and $\mathcal{H}$ is one of the standard choices in Definition 2.34, there are a number of known convenient choices of $\mathcal{Y}$.

In the next two sections, we will present characterizing operators for the normal and Poisson distributions, and some associated convenient choices for $\mathcal{Y}$.

4 The Normal Case

We'll use Proposition 3.3 to give our first example of a characterizing operator.

Theorem 4.1. *Let $X_{\mathcal{N}} \in \mathcal{N}(0,1)$ and let*

$$\mathcal{V}_{\mathcal{N}} = \Big\{ h \text{ absolutely continuous} : \mathbb{E}\Big[|X_{\mathcal{N}}|^k |h(X_{\mathcal{N}})| \Big] < \infty \text{ for all } k \geq 0 \Big\}.$$

Let $\mathcal{F}_{\mathcal{N}}$ be the set of absolutely continuous f with $f' \in \mathcal{V}_{\mathcal{N}}$. (For a definition and characterization of absolute continuity, see [Rud66, Definition 7.17 and Theorem 7.18]).

The linear operator $T_{\mathcal{N}} : \mathcal{F}_{\mathcal{N}} \to \mathcal{V}_{\mathcal{N}}$ given by $T_{\mathcal{N}} f(x) = f'(x) - xf(x)$ is a characterizing operator for $\mathcal{N}(0,1)$.

Proof. (Adapted from [Ste86, Lecture II]).

First, for each $k \geq 0$ we have $e^{x^2/4} > x^k$ for sufficiently large x, so

$$\int_C^\infty x^k e^{-x^2/2} \, \mathrm{d}\, x < \int_C^\infty e^{-x^2/4} \, \mathrm{d}\, x < \infty$$

for some large C. It follows that $\mathbb{E}|X_{\mathcal{N}}|^{k+1} < \infty$, and $\mathcal{V}_{\mathcal{N}}$ contains the constant functions.

Now, we prove that $T_{\mathcal{N}}$ is well-defined as an operator with codomain $\mathcal{V}_{\mathcal{N}}$. Fix $f \in \mathcal{F}_{\mathcal{N}}$ and $k \geq 0$. We have

$$\begin{aligned}
&\mathbb{E}\Big[|X_{\mathcal{N}}|^k |T_{\mathcal{N}} f(X_{\mathcal{N}})|\Big] \\
&\quad \leq \mathbb{E}\Big[|X_{\mathcal{N}}|^k \big|f'(X_{\mathcal{N}})\big|\Big] + \mathbb{E}\Big[|X_{\mathcal{N}}|^{k+1} |f(X_{\mathcal{N}}) - f(0)|\Big] + |f(0)| \mathbb{E}\Big[|X_{\mathcal{N}}|^{k+1}\Big]. \qquad (4.1)
\end{aligned}$$

Comment 4.1. Stein had $\int_t^\infty x^k e^{-x^2/2} \,\mathrm{d}\, x \leq C_3\big(1 + t^k\big)$, but I couldn't immediately see how to prove that, and this works too.

The first term in (4.1) is finite by the definition of $\mathcal{F}_{\mathcal{N}}$, and the third term is finite, as above. We will now bound the second term. Note that

$$\begin{aligned}
\int_t^\infty x^{k+1} e^{-x^2/2} \,\mathrm{d}\, x &\leq e^{-t^2/2} \int_0^\infty (y+t)^{k+1} e^{-y^2/2} \,\mathrm{d}\, y \\
&= O\Big(e^{-t^2/2}\Big(1 + t + \ldots t^{k+1}\Big)\Big) \\
&= O\Big(e^{-t^2/2}\Big(1 + t^{k+1}\Big)\Big),
\end{aligned}$$

with the substitution $y = x - t$. With Fubini's theorem (see [Rud66, Theorem 8.8]),

$$\begin{aligned}
\mathbb{E}\Big[|X_{\mathcal{N}}|^{k+1}|f(X_{\mathcal{N}})|\Big] &= O\bigg(\int_0^\infty x^{k+1}\bigg|\int_0^x f'(t)\,\mathrm{d}\,t\bigg|e^{-x^2/2}\,\mathrm{d}\,x \\
&\qquad - \int_{-\infty}^0 x^{k+1}\bigg|\int_0^x f'(t)\,\mathrm{d}\,t\bigg|e^{-x^2/2}\,\mathrm{d}\,x\bigg) \\
&= O\bigg(\int_0^\infty \int_t^\infty x^{k+1}\big|f'(t)\big|e^{-x^2/2}\,\mathrm{d}\,x\,\mathrm{d}\,t \\
&\qquad - \int_{-\infty}^0 \int_{-\infty}^t x^{k+1}\big|f'(t)\big|e^{-x^2/2}\,\mathrm{d}\,x\,\mathrm{d}\,t\bigg) \\
&= O\bigg(\int_0^\infty e^{-t^2/2}\Big(1+t^{k+1}\Big)\big|f'(t)\big|\,\mathrm{d}\,t \\
&\qquad - \int_{-\infty}^0 e^{-t^2/2}\Big(1+t^{k+1}\Big)\big|f'(t)\big|\,\mathrm{d}\,t.\bigg) \\
&= O\Big(\mathbb{E}\Big[|X_{\mathcal{N}}|^0\big|f'(X_{\mathcal{N}})\big|\Big] + \mathbb{E}\Big[|X_{\mathcal{N}}|^{k+1}\big|f'(X_{\mathcal{N}})\big|\Big]\Big) \\
&< \infty.
\end{aligned}$$

It follows that $\mathbb{E}\Big[|X_{\mathcal{N}}|^k|T_{\mathcal{N}}f(X_{\mathcal{N}})|\Big] < \infty$ and $T_{\mathcal{N}}f \in \mathcal{V}_{\mathcal{N}}$.

Now, for any $f \in \mathcal{F}_{\mathcal{N}}$, integration by parts gives

$$\mathbb{E}_{\mathcal{N}}T_{\mathcal{N}}f = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty e^{-t^2/2}f'(t)\,\mathrm{d}\,t - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty te^{-t^2/2}f(t)\,\mathrm{d}\,t = 0 \tag{4.2}$$

so $\mathbb{E}_{\mathcal{N}}T_{\mathcal{N}} = 0$ and (3.1) holds.

Next, consider the differential equation

$$T_{\mathcal{N}}f(x) = f'(x) - xf(x) = h(x) - \mathbb{E}_{\mathcal{N}}h \tag{4.3}$$

for f. We solve this equation with the method of integrating factors:

$$\begin{aligned}
e^{-x^2/2}f'(x) - xe^{-x^2/2}f(x) &= e^{-x^2/2}(h(x) - \mathbb{E}_{\mathcal{N}}h), \\
\frac{\mathrm{d}}{\mathrm{d}\,x}e^{-x^2/2}f(x) &= e^{-x^2/2}(h(x) - \mathbb{E}_{\mathcal{N}}h).
\end{aligned}$$

The general solution is

$$f(x) = e^{x^2/2} \int_{-\infty}^{x} (h(t) - \mathbb{E}_{\mathcal{N}} h) e^{-t^2/2} \,\mathrm{d}\, t + c e^{x^2/2}. \tag{4.4}$$

We can therefore define $U_{\mathcal{N}} : \mathcal{V}_{\mathcal{N}} \to \mathcal{F}_{\mathcal{N}}$ by

$$U_{\mathcal{N}} h(x) = e^{x^2/2} \int_{-\infty}^{x} (h(t) - \mathbb{E}_{\mathcal{N}} h) e^{-t^2/2} \,\mathrm{d}\, t = -e^{x^2/2} \int_{x}^{\infty} (h(t) - \mathbb{E}_{\mathcal{N}} h) e^{-t^2/2} \,\mathrm{d}\, t. \tag{4.5}$$

(These two definitions are the same because their difference is $e^{x^2/2}(\mathbb{E}_{\mathcal{N}} h - \mathbb{E}_{\mathcal{N}} h) = 0$). By construction,

$$T_{\mathcal{N}} U_{\mathcal{N}} h(x) = h(x) - \mathbb{E}_{\mathcal{N}} h.$$

So, (3.2) holds. To apply Proposition 3.3, it remains to prove that $U_{\mathcal{N}}$ is indeed well-defined as an operator with codomain $\mathcal{F}_{\mathcal{N}}$. Since $U_{\mathcal{N}} h$ satisfies (4.3), we have

$$\begin{aligned} \mathbb{E}\Big[|X_{\mathcal{N}}|^k \big|(U_{\mathcal{N}} h)'(X_{\mathcal{N}})\big|\Big] &= \mathbb{E}\Big[|X_{\mathcal{N}}|^k |X_{\mathcal{N}} U_{\mathcal{N}} h(X_{\mathcal{N}}) + h(X_{\mathcal{N}}) - \mathbb{E}_{\mathcal{N}} h|\Big] \\ &\le \mathbb{E}\Big[|X_{\mathcal{N}}|^{k+1} |U_{\mathcal{N}} h(X_{\mathcal{N}})|\Big] + \mathbb{E}\Big[|X_{\mathcal{N}}|^k |h(X_{\mathcal{N}})|\Big] + \mathbb{E}|X_{\mathcal{N}}|^k \mathbb{E}_{\mathcal{N}} h. \end{aligned} \tag{4.6}$$

for $h \in \mathcal{V}_{\mathcal{N}}$. The first and third terms of (4.6) are finite by the definition of $\mathcal{V}_{\mathcal{N}}$ and the fact that it contains the constant function 1. Finally, for the second term, Fubini's theorem gives

$$\begin{aligned} \mathbb{E}\Big[|X_{\mathcal{N}}|^{k+1} |U_{\mathcal{N}} h(X_{\mathcal{N}})|\Big] &= O\bigg(\int_0^\infty x^{k+1} \int_x^\infty |h(t) - \mathbb{E}_{\mathcal{N}} h| e^{-t^2/2} \,\mathrm{d}\, t \,\mathrm{d}\, x \\ &\qquad - \int_{-\infty}^0 x^{k+1} \int_\infty^x |h(t) - \mathbb{E}_{\mathcal{N}} h| e^{-t^2/2} \,\mathrm{d}\, t \,\mathrm{d}\, x \bigg) \\ &= O\bigg(\int_0^\infty |h(t) - \mathbb{E}_{\mathcal{N}} h| e^{-t^2/2} \int_0^t x^{k+1} \,\mathrm{d}\, x \,\mathrm{d}\, t \\ &\qquad - \int_{-\infty}^0 |h(t) - \mathbb{E}_{\mathcal{N}} h| e^{-t^2/2} \int_t^0 x^{k+1} \,\mathrm{d}\, x \,\mathrm{d}\, t \bigg) \\ &= O\bigg(\int_{-\infty}^\infty |t|^{k+2} |h(t) - \mathbb{E}_{\mathcal{N}} h| e^{-t^2/2} \,\mathrm{d}\, x \,\mathrm{d}\, t \bigg) \\ &= O\Big(\mathbb{E}\Big[|X_{\mathcal{N}}|^{k+2} |h(X_{\mathcal{N}})|\Big] + \mathbb{E}\Big[|X_{\mathcal{N}}|^{k+2}\Big] \Big) < \infty. \end{aligned}$$

That is, $\mathbb{E}\Big[|X_{\mathcal{N}}|^k\big|(U_{\mathcal{N}}h)'(X_{\mathcal{N}})\big|\Big] < \infty$, so $U_{\mathcal{N}}h \in \mathcal{F}_{\mathcal{N}}$. □

Remark 4.2. Note that $\mathcal{H}_{\mathrm{W}} \subseteq \mathcal{V}_{\mathcal{N}}$, where $\mathcal{H}_{\mathrm{W}}$ is as defined in Definition 2.34. Since $\mathcal{H}_{\mathrm{W}}$ is a determining class, $T_{\mathcal{N}}$ is a characterizing operator in the sense of Proposition 3.2.

Useful application of Stein's method depends on a good understanding of the functions in $U_{\mathcal{N}}\mathcal{H}$. This will give us a set of functions $\mathcal{Y}$ to maximize over, as in (3.4). In the case of the Wasserstein and bounded Lipschitz metric, we can show that the first three derivatives are all bounded, so we can choose $\mathcal{Y}$ to be the set of functions that satisfies these bounds. This makes these metrics good choices for normal approximation.

Theorem 4.3. *Suppose that* $\mathcal{H} \in \{\mathcal{H}_{\mathrm{W}}, \mathcal{H}_{\mathrm{BL}}\}$ *and* $\mathcal{L}_0 = \mathcal{N}(0,1)$. *Using the characterizing operator in Theorem 4.1,*

$$\|U_{\mathcal{N}}h\|_\infty \leq 2, \quad \big\|(U_{\mathcal{N}}h)'\big\|_\infty \leq 3\sqrt{\pi/2}, \quad \big\|(U_{\mathcal{N}}h)''\big\|_\infty \leq 6.$$

Proof. (Adapted from [Ste12, Section 3.1]).

First, note that Lipschitz functions are trivially absolutely continuous, and

$$\mathbb{E}\Big[|X_{\mathcal{N}}|^k|h(X_{\mathcal{N}})|\Big] < |h(0)|\mathbb{E}|X_{\mathcal{N}}|^k + \mathbb{E}|X_{\mathcal{N}}|^{k+1}$$

for $h \in \mathcal{H}$. So, $\mathcal{H} \subseteq \mathcal{V}_{\mathcal{N}}$ and it makes sense to consider $U_{\mathcal{N}}\mathcal{H}$.

Since $U_{\mathcal{N}}h = U_{\mathcal{N}}(h - h(0))$ for all $h \in \mathcal{H}_{\mathcal{N}}$, it suffices to consider $h \in \mathcal{H}$ with $h(0) = 0$. So, fix such a $h \in \mathcal{H}$. We have $h(x) \leq |x|$ for all $x \in \mathbb{R}$, and it follows that

$$|\mathbb{E}_{\mathcal{N}}h| \leq \mathbb{E}|X_{\mathcal{N}}| = \frac{2}{\sqrt{2\pi}}\int_0^\infty x e^{-x^2/2}\,\mathrm{d}\,x = \sqrt{2/\pi}.$$

Then, for $x \geq 0$,

$$\begin{aligned}
|U_{\mathcal{N}}h(x)| &\leq e^{x^2/2}\int_x^\infty |h(t) - \mathbb{E}_{\mathcal{N}}h|e^{-t^2/2}\,\mathrm{d}\,t \\
&\leq e^{x^2/2}\int_x^\infty \left(t + \sqrt{\frac{2}{\pi}}\right)e^{-t^2/2}\,\mathrm{d}\,t \\
&\leq e^{x^2/2}\int_x^\infty te^{-t^2/2}\,\mathrm{d}\,t + \sqrt{\frac{2}{\pi}}\int_0^\infty e^{-s^2/2}\,\mathrm{d}\,s \\
&= 2.
\end{aligned}$$

(We have used the substitution $s = t - x$, and the Gaussian integral). With the alternative form for $U_{\mathcal{N}}$ in (4.5), an identical argument gives $|U_{\mathcal{N}}h(x)| \leq 2$ for $x \leq 0$, so $\|U_{\mathcal{N}}h\|_\infty \leq 2$.

Now, note that

$$(T_{\mathcal{N}}U_{\mathcal{N}}h)'(x) = (U_{\mathcal{N}}h)''(x) - x(U_{\mathcal{N}}h)'(x) - U_{\mathcal{N}}h(x) = h'(x).$$

Set $g = h' + U_{\mathcal{N}}h$. Note $\|g\|_\infty \leq \|h'\|_\infty + \|U_{\mathcal{N}}h\|_\infty \leq 3$, so $g \in \mathcal{V}_{\mathcal{N}}$, and

$$\mathbb{E}_{\mathcal{N}}g = \mathbb{E}\big[(U_{\mathcal{N}}h)''(X_{\mathcal{N}}) - X_{\mathcal{N}}(U_{\mathcal{N}}h)'(X_{\mathcal{N}})\big] = 0$$

as in (4.2). Hence

$$T_{\mathcal{N}}\big((U_{\mathcal{N}}h)'\big) = g + \mathbb{E}_{\mathcal{N}}g,$$

and $(U_{\mathcal{N}}h)'$ satisfies (4.3) (with g taking the role of h). By the general form (4.4) of the solution to (4.3), we must have

$$(U_{\mathcal{N}}h)' = e^{x^2/2}\int_{-\infty}^x (h(t) - \mathbb{E}_{\mathcal{N}}h)e^{-t^2/2}\,\mathrm{d}\,t + ce^{x^2/2}$$

for some c. Since $(U_{\mathcal{N}}h)' \in \mathcal{V}_{\mathcal{N}}$, we must have $e^{-x^2/2}(U_{\mathcal{N}}h)'(x) \to 0$. That is, $c = 0$. For $x \geq 0$, we have

$$\big|(U_{\mathcal{N}}h)'(x)\big| \leq e^{x^2/2}\int_x^\infty |g(t)|e^{-t^2/2}\,\mathrm{d}\,t \leq 3\int_0^\infty e^{-s^2/2}\,\mathrm{d}\,s \leq 3\sqrt{\frac{\pi}{2}}.$$

(As before, $s = t - x$). A similar calculation shows that $\left|(U_{\mathcal{N}}h)'(x)\right| \leq 3\sqrt{\frac{\pi}{2}}$ for $x < 0$, so

$$\left\|(U_{\mathcal{N}}h)'\right\|_\infty \leq 3\sqrt{\frac{\pi}{2}}.$$

Also, for $x \geq 0$,

$$\begin{aligned}
\left|(U_{\mathcal{N}}g)'(x)\right| &= |g(x) + xU_{\mathcal{N}}g(x)| \\
&\leq 3 + 3xe^{x^2/2}\int_x^\infty e^{-t^2/2}\,\mathrm{d}\,t \\
&\leq 3 + 3xe^{x^2/2}\int_x^\infty \frac{t}{x}e^{-t^2/2}\,\mathrm{d}\,t \\
&= 6.
\end{aligned}$$

Similarly, $\left|(U_{\mathcal{N}}g)'(x)\right| \leq 6$ for $x < 0$, so

$$\left\|(U_{\mathcal{N}}h)''\right\|_\infty = \left\|(U_{\mathcal{N}}g)'\right\|_\infty \leq 6.$$

□

Remark 4.4. In the interests of a simple proof, the bounds in Theorem 4.3 are very crude. The stronger bounds $\|f'\|_\infty \leq \sqrt{\frac{2}{\pi}}$ and $\|f''\|_\infty \leq 2$ can be shown to hold, with similarly elementary but much more fiddly proofs. See [CGS10, Lemma 2.4].

4.1 The Berry-Esseen Theorem

The Berry-Esseen Theorem is a quantitative version of the Central Limit Theorem. The original version was independently proved in [Ber41] and [Ess42] with some relatively involved Fourier analysis. However, the machinery of Stein's method allows for a very simple proof.

The purpose of this section is to make the theory from Chapter 3 a bit more concrete by giving our first example of Stein's method in action. Of course, the result is also of practical interest in statistics.

Theorem 4.5. *Let $Q_1, \ldots, Q_n \in \mathcal{L}_Q$ be independent, with $\mathbb{E}Q = 0$, $\mathbb{E}Q^2 = 1$ and $\mathbb{E}|Q|^3 < \infty$.*

Define

$$X = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Q_i.$$

Then

$$d_{\mathrm{W}}(X, \mathcal{N}(0,1)) \leq \frac{C}{\sqrt{n}} \mathbb{E}|Q|^3$$

for some constant C.

Remark 4.6. In the following proof, we will prove the theoren with $C = 9$. The sharper bounds in Remark 4.4 yield $C = 3$ with the same argument.

Remark 4.7. Proposition 2.40 allows us to transfer the result of Theorem 4.5 from the Wasserstein metric to the Kolmogorov metric. The Berry-Esseen Theorem is usually stated for the Kolmogorov metric, because it has a direct interpretation as the supremum difference between probabilities. However in this case the transfer makes the bound much worse. It is possible to use Stein's method directly in the Kolmogorov metric (see [Ste12, Section 6]), but this is a little more involved.

Proof of Theorem 4.5. (Adapted from [Ste12, Section 5]). Let f be a function satisfying the bounds in Theorem 4.3

We need to compare

$$\mathbb{E}[Xf(X)] = \mathbb{E}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Q_i f(X)\right] \tag{4.7}$$

with $\mathbb{E}f'(X)$. Define $W_i = X - \frac{1}{\sqrt{n}} Q_i$, so each W_i and Q_i are independent. For each i we then have $\mathbb{E}Q_i f(W_i) = \mathbb{E}Q_i \mathbb{E}f(W_i) = 0$, so

$$\mathbb{E}[Q_i f(X)] = \mathbb{E}[Q_i(f(X) - f(W_i))]. \tag{4.8}$$

Now, we use a Taylor expansion with the Lagrange form of the remainder. Noting that $X - W_i = \frac{1}{\sqrt{n}} Q_i$, we have

$$f(X) = f(W_i) + \frac{1}{\sqrt{n}} Q_i f'(W_i) + \frac{1}{2n} Q_i^2 f''(E)$$

for some (random) E. Combining this with (4.8) gives

$$\mathbb{E}[Q_i f(X)] = \frac{1}{\sqrt{n}}\mathbb{E}\big[Q_i^2 f'(W_i)\big] + \frac{1}{2n}\mathbb{E}\big[Q_i^3 f''(E)\big]. \tag{4.9}$$

Next, note that

$$\mathbb{E}\big[Q_i^2 f'(W_i)\big] = \mathbb{E}Q_i^2\,\mathbb{E}f'(W_i) = \mathbb{E}f'(W_i) = \mathbb{E}f'(X) + \frac{1}{\sqrt{n}}\,\mathbb{E}\big[Q_i f''\big(E'\big)\big] \tag{4.10}$$

for some E', again using the Lagrange form of the remainder. Since $\|f''\|_\infty \le 6$, we have

$$\frac{1}{2}\mathbb{E}\big|Q_i^3 f''(E)\big| \le 3\,\mathbb{E}\big|Q^3\big|, \quad \mathbb{E}\big|Q_i f''\big(E'\big)\big| \le 6\,\mathbb{E}|Q|. \tag{4.11}$$

Hölder's inequality (see [Rud66, Theorem 3.6 (1)]) for the $L^{3/2}$ norm gives

$$1 = \mathbb{E}\big|Q^2\big|^{3/2} \le \left(\mathbb{E}\Big|\big(Q^2\big)^{3/2}\Big|^{2/3}\right)^{3/2} = \mathbb{E}\big|Q^3\big|,$$

so an application of Hölder's inequality for the L^3 norm gives

$$\mathbb{E}|Q| \le \mathbb{E}\big|Q^3\big|^{1/3} \le \mathbb{E}\big|Q^3\big|. \tag{4.12}$$

Combining Equations (4.7) and (4.9) to (4.12) gives

$$\begin{aligned}
\big|\mathbb{E}[Xf(X)] - \mathbb{E}f'(X)\big| &\le \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\mathbb{E}[Q_i f(X)] - \frac{1}{\sqrt{n}}\mathbb{E}\big[Q_i^2 f'(W_i)\big]\right)\right| \\
&\qquad + \left|\frac{1}{n}\sum_{i=1}^n\big(\mathbb{E}\big[Q_i^2 f'(W_i)\big] - \mathbb{E}f'(X)\big)\right| \\
&\le \frac{3}{\sqrt{n}}\,\mathbb{E}\big|Q^3\big| + \frac{6}{\sqrt{n}}\,\mathbb{E}\big|Q^3\big| \\
&= \frac{9}{\sqrt{n}}\,\mathbb{E}\big|Q^3\big|.
\end{aligned}$$

□

5 The Poisson Case

Theorem 5.1. *Let* $X_{\mathrm{Po}(\lambda)} \in \mathrm{Po}(\lambda)$ *and let*

$$\mathcal{V}_{\mathrm{Po}(\lambda)} = \mathcal{F}_{\mathrm{Po}(\lambda)} = \Big\{h : \mathbb{N} \to \mathbb{R} : \mathbb{E}\Big[X^k_{\mathrm{Po}(\lambda)} \big| h\big(X_{\mathrm{Po}(\lambda)}\big)\big|\Big] < \infty \text{ for all } k \in \mathbb{N}\Big\}.$$

The operator $T_{\mathrm{Po}(\lambda)} : \mathcal{F}_{\mathrm{Po}(\lambda)} \to \mathcal{V}_{\mathrm{Po}(\lambda)}$ *defined by* $T_{\mathrm{Po}(\lambda)} f(k) = \lambda f(k+1) - k f(k)$ *is a characterizing operator for* $\mathrm{Po}(\lambda)$.

Proof. The components of the proof are very similar to that of Theorem 4.1. First, fix k and note that

$$(i+1)^k \frac{\lambda^{(i+1)}}{(i+1)!} \Big/ i^k \frac{\lambda^i}{i!} \to 0$$

as $i \to \infty$. So by the ratio test $\mathbb{E}X^k_{\mathrm{Po}(\lambda)} = O\Big(\sum_{i=0}^{\infty} i^k \frac{\lambda^i}{i!}\Big) < \infty$, and $\mathcal{V}_{\mathcal{N}}$ contains the constant functions.

Now, we check that $T_{\mathcal{N}} f$ is well-defined as a linear operator with codomain $\mathcal{V}_{\mathrm{Po}(\lambda)}$. For $f \in \mathcal{F}_{\mathrm{Po}(\lambda)}$ and $k \in \mathbb{N}$,

$$\mathbb{E}\Big[X^k_{\mathrm{Po}(\lambda)} \big| T_{\mathrm{Po}(\lambda)} f\big(X_{\mathrm{Po}(\lambda)}\big)\big|\Big] \leq \lambda \mathbb{E}\Big[X^k_{\mathrm{Po}(\lambda)} \big| f\big(X_{\mathrm{Po}(\lambda)} + 1\big)\big|\Big] + \mathbb{E}\Big[X^{k+1}_{\mathrm{Po}(\lambda)} \big| f\big(X_{\mathrm{Po}(\lambda)}\big)\big|\Big].$$

We have

$$\begin{aligned}
\mathbb{E}\Big[X^k_{\mathrm{Po}(\lambda)} \big| f\big(X_{\mathrm{Po}(\lambda)} + 1\big)\big|\Big] &= O\left(\sum_{i=0}^{\infty} i^k \frac{\lambda^i}{i!} |f(i+1)|\right) \\
&= O\left(\sum_{r=0}^{k+1} \sum_{i=0}^{\infty} (i+1)^r \frac{\lambda^{i+1}}{(i+1)!} |f(i+1)|\right) \\
&= O\left(\sum_{r=0}^{k+1} \mathbb{E}\Big[X^k_{\mathrm{Po}(\lambda)} \big| f\big(X_{\mathrm{Po}(\lambda)}\big)\big|\Big]\right) \\
&< \infty.
\end{aligned}$$

It follows that $\mathbb{E}\Big[X^k_{\mathrm{Po}(\lambda)} \big| T_{\mathrm{Po}(\lambda)} f\big(X_{\mathrm{Po}(\lambda)}\big)\big|\Big] < \infty$ and $T_{\mathrm{Po}(\lambda)} f \in \mathcal{V}_{\mathrm{Po}(\lambda)}$.

For any $f \in \mathcal{F}_{\mathrm{Po}(\lambda)}$, we have

$$\mathbb{E}_{\mathrm{Po}(\lambda)} T_{\mathrm{Po}(\lambda)} f = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^{i+1}}{i!} f(i+1) - e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^i}{(i-1)!} f(i) = 0$$

so $\mathbb{E}_{\mathrm{Po}(\lambda)} T_{\mathrm{Po}(\lambda)} = 0$ and (3.1) holds. Then, define $U_{\mathrm{Po}(\lambda)} : \mathcal{V}_{\mathrm{Po}(\lambda)} \to \mathcal{F}_{\mathrm{Po}(\lambda)}$ by

$$U_{\mathrm{Po}(\lambda)} h(i) = \frac{(i-1)!}{\lambda^i} \sum_{j=0}^{i-1} \frac{\lambda^j}{j!} \big(h(j) - \mathbb{E}_{\mathrm{Po}(\lambda)} h \big)$$

for $i \in \mathbb{Z}^+$ (and $U_{\mathrm{Po}(\lambda)} h(0) = 0$).

For $h \in \mathcal{V}_{\mathrm{Po}(\lambda)}$ and $k \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}\Big[X_{\mathrm{Po}(\lambda)}^k \big| U_{\mathrm{Po}(\lambda)} h\big(X_{\mathrm{Po}(\lambda)}\big) \big| \Big] &= O\left(\sum_{i=1}^{\infty} i^{k-1} \left| \sum_{j=0}^{i-1} \frac{\lambda^j}{j!} \big(h(j) - \mathbb{E}_{\mathrm{Po}(\lambda)} h \big) \right| \right) \\
&= O\left(\sum_{i=1}^{\infty} i^{k-1} \left| \sum_{j=i}^{\infty} \frac{\lambda^j}{j!} \big(h(j) - \mathbb{E}_{\mathrm{Po}(\lambda)} h \big) \right| \right) \\
&= O\left(\sum_{j=1}^{\infty} \frac{\lambda^j}{j!} \big| h(j) - \mathbb{E}_{\mathrm{Po}(\lambda)} h \big| \sum_{i=1}^{j} i^{k-1} \right) \\
&= O\left(\sum_{j=1}^{\infty} j^k \frac{\lambda^j}{j!} \big| h(j) - \mathbb{E}_{\mathrm{Po}(\lambda)} h \big| \right) \\
&< \infty,
\end{aligned}$$

so $U_{\mathrm{Po}(\lambda)} h \in \mathcal{F}_{\mathrm{Po}(\lambda)}$, and $U_{\mathrm{Po}(\lambda)}$ is well-defined with codomain $\mathcal{F}_{\mathrm{Po}(\lambda)}$. Now, substituting and simplifying gives

$$T_{\mathrm{Po}(\lambda)} U_{\mathrm{Po}(\lambda)} h(i) = h(i) - \mathbb{E}_{\mathrm{Po}(\lambda)} h,$$

so (3.2) holds and Proposition 3.3 completes the proof. □

Remark 5.2. As for the normal case, $T_{\mathrm{Po}(\lambda)}$ is a characterizing operator in the sense of Proposition 3.2, because $\mathcal{H}_{\mathrm{TV}} \subseteq \mathcal{V}_{\mathrm{Po}(\lambda)}$.

The natural metric to use in Poisson approximation is the total variation metric. We can bound the functions in $U_{\mathrm{Po}(\lambda)} \mathcal{H}_{\mathrm{TV}}$ as well as their first differences (define $\Delta f(i) = f(i+1) - f(i)$).

Theorem 5.3. *Suppose $\mathcal{H} = \mathcal{H}_{\mathrm{TV}}$ and $\mathcal{L}_0 = \mathrm{Po}(\lambda)$. Using the characterizing operator in Theorem 5.1,*

$$\left\|U_{\mathrm{Po}(\lambda)}h\right\|_\infty \leq \min\left\{2, 3\lambda^{-1/2}\right\}, \quad \left\|\Delta U_{\mathrm{Po}(\lambda)}h\right\|_\infty \leq \lambda^{-1}\left(1 - e^{-\lambda}\right)$$

for all $h \in \mathcal{H}$.

Proof. (Adapted from [BHJ92, Lemma 1.1.1]). For any Borel A and any $k \in \mathbb{N}$,

$$\begin{aligned}
U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_A(i) &= \frac{e^\lambda\,(i-1)!}{\lambda^i}(\mathrm{Po}_\lambda(A \cap [0, i-1]) - \mathrm{Po}_\lambda(A)\,\mathrm{Po}_\lambda([0, i-1])) \\
&= \frac{e^\lambda\,(i-1)!}{\lambda^i}\Big(\,\mathrm{Po}_\lambda(A \cap [0, i-1])\Big(1 - \mathrm{Po}_\lambda([0, i-1])\Big) \\
&\qquad\qquad - \Big(\,\mathrm{Po}_\lambda(A) - \mathrm{Po}_\lambda(A \cap [0, i-1])\Big)\,\mathrm{Po}_\lambda([0, i-1])\Big) \\
&= \frac{e^\lambda\,(i-1)!}{\lambda^i}\Big(\,\mathrm{Po}_\lambda(A \cap [0, i-1])\,\mathrm{Po}_\lambda([i, \infty]) - \mathrm{Po}_\lambda(A \backslash [0, i-1])\,\mathrm{Po}_\lambda([0, i-1])\Big).
\end{aligned} \tag{5.1}$$

Note that $\mathrm{Po}_\lambda(A \cap [0, i-1])$ is bounded above by $\mathrm{Po}_\lambda([0, i-1])$ and $\mathrm{Po}_\lambda(A \backslash [0, i-1])$ is bounded above by $\mathrm{Po}_\lambda([i, \infty])$, so

$$\left|U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_A(i)\right| \leq \frac{e^\lambda\,(i-1)!}{\lambda^i}\,\mathrm{Po}_\lambda([0, i-1])\,\mathrm{Po}_\lambda([i, \infty]).$$

If $1 \leq i \leq \lambda$, then

$$\begin{aligned}
\left|U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_A(i)\right| &\leq \frac{e^\lambda\,(i-1)!}{\lambda^i}\,\mathrm{Po}_\lambda([0, i-1]) \\
&= \lambda^{-1}\sum_{j=0}^{i-1}\frac{(i-1)!}{j!\lambda^{i-1-j}} \\
&= \lambda^{-1}\sum_{q=0}^{i-1}\frac{(i-1)!}{(i-1-q)!\lambda^q}.
\end{aligned}$$

Now, for $q = 0, \ldots, i-1$ define

$$a_q = \frac{(i-1)!}{(i-1-q)!\lambda^q}.$$

Note $a_q \le (i-1)^q/\lambda^q \le 1$ for all q, so immediately $\left|U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_A(i)\right| \le i/\lambda \le 1$. Also,

$$a_q \le \frac{i-q}{\lambda} \le 1-\lambda^{-1/2}, \text{ and } a_{q+1}/a_q = \frac{i-q}{\lambda} \le 1-\lambda^{-1/2}$$

for $q > \lambda^{1/2}$. So,

$$\begin{aligned}
\left|U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_A(i)\right| &\le \lambda^{-1}\sum_{q=0}^{\lfloor\lambda^{1/2}\rfloor} a_q + \lambda^{-1}\sum_{q=\lceil\lambda^{1/2}\rceil}^{i-1} a_q \\
&\le \lambda^{-1}\Big(1+\lambda^{1/2}\Big) + \lambda^{-1}\sum_{r=1}^{\infty}\Big(1-\lambda^{-1/2}\Big)^r \\
&= 2\lambda^{-1/2}
\end{aligned}$$

(note $\lambda \ge i \ge 1$, so $0 \le 1-\lambda^{-1/2} < 1$). Now, suppose that $i > \lambda$. Then

$$\begin{aligned}
\left|U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_A(i)\right| &\le \frac{e^\lambda\,(i-1)!}{\lambda^i}\,\mathrm{Po}_\lambda([i,\infty]) \\
&= i^{-1}\sum_{j=i}^{\infty}\frac{i!}{j!\lambda^{i-j}} \\
&= i^{-1}\sum_{q=0}^{\infty}\frac{\lambda^q i!}{(i+q)!} \\
&\le i^{-1}\sum_{j=i}^{\infty}\frac{\lambda^q}{(i+1)^q} \\
&= \frac{i+1}{i(i+1-\lambda)} \\
&\le 2.
\end{aligned}$$

This also shows that $\left|U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_A(i)\right| \le 2\lambda^{-1/2}$ if $\lambda < 1$. Now, let $b_q = \lambda^q i!/(i+q)!$. As for the previous case,

$$b_q \le \frac{\lambda^q}{(i+1)^q} \le 1.$$

Also, for $q > \lambda^{1/2}$,

$$b_q \le \frac{\lambda}{i+q} < \frac{\lambda}{\lambda+\lambda^{1/2}} \text{ and } \frac{b_{q+1}}{b_q} \le \frac{\lambda}{i+q} < \frac{\lambda}{\lambda+\lambda^{1/2}}.$$

So,

$$\begin{aligned}
\left|U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_A(i)\right| &= i^{-1}\sum_{q=0}^{\lfloor\lambda^{1/2}\rfloor} b_q + i^{-1}\sum_{q=\lceil\lambda^{1/2}\rceil}^{\infty} b_q \\
&= \lambda^{-1}\Big(1+\lambda^{1/2}\Big) + \lambda^{-1}\sum_{r=1}^{\infty}\bigg(\frac{\lambda}{\lambda+\lambda^{1/2}}\bigg)^r \qquad (5.2)\\
&= \lambda^{-1} + 2\lambda^{-1/2} \\
&\leq 3\lambda^{-1/2}.
\end{aligned}$$

We have proved that $\left\|U_{\mathrm{Po}(\lambda)}h\right\|_\infty \leq \min\left\{2, 3\lambda^{-1/2}\right\}$ for $h \in \mathcal{H}$.

Comment 5.1. [BHJ92, Lemma 1.1.1] gives this argument (but with no details), and claims it yields $\left|U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_A(i)\right| \leq 2\lambda^{-1/2}$. I'm not seeing it though... I might revisit this.

Let $i, k \in \mathbb{N}$. By (5.1),

$$\begin{aligned}
U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_{\{k\}}(i-1) &= -\frac{\lambda^k}{e^\lambda k!}\sum_{j=0}^{i-2}\lambda^{j-i}\frac{(i-1)!}{j!} \\
&= -\frac{\lambda^k}{e^\lambda k!}\sum_{q=1}^{i-1}\lambda^{q-i}\frac{(i-2)!}{(q-1)!} \\
&\leq -\frac{\lambda^k}{e^\lambda k!}\sum_{q=0}^{i-1}\lambda^{q-i}\frac{(i-1)!}{q!} \\
&= U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_{\{k\}}(i).
\end{aligned}$$

Similarly, if $i > k$ then

$$\begin{aligned} U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_{\{k\}}(i) &= \frac{\lambda^k}{e^\lambda k!}\sum_{j=i}^{\infty}\lambda^{j-i}\frac{(i-1)!}{j!} \\ &= \frac{\lambda^k}{e^\lambda k!}\sum_{q=i+1}^{\infty}\lambda^{q-i-1}\frac{(i-1)!}{(q-1)!} \\ &\leq \frac{\lambda^k}{e^\lambda k!}\sum_{q=i+1}^{\infty}\lambda^{q-i-1}\frac{i!}{q!} \\ &= U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_{\{k\}}(i+1). \end{aligned}$$

Therefore,

$$U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_{\{k\}}(i+1) - U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_{\{k\}}(i)$$

is positive only when $i = k$, and

$$\begin{aligned} U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_{\{k\}}(k+1) - U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_{\{k\}}(k) &= e^{-\lambda}\lambda^{-1}\sum_{j=k+1}^{\infty}\frac{\lambda^j}{j!} + e^{-\lambda}\lambda^{-1}\sum_{r=1}^{k}\frac{\lambda^r}{r!}\frac{r}{k} \\ &\leq e^{-\lambda}\lambda^{-1}\Big(e^\lambda - 1\Big) = \lambda^{-1}\Big(1 - e^{-\lambda}\Big). \end{aligned}$$

Now, for any $A \subseteq \mathbb{N}$, $\mathbf{1}_A = \sum_{k\in A}\mathbf{1}_{\{k\}}$, so linearity of $U_{\mathrm{Po}(\lambda)}$ gives

$$\Delta U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_A(k) \leq \lambda^{-1}\Big(1 - e^{-\lambda}\Big)$$

for all k. Finally, note that $U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_{\mathbb{N}\setminus A} + U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_A = U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_{\mathbb{N}} = 0$, so

$$-\Delta U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_A(k) \geq \Delta U_{\mathrm{Po}(\lambda)}\,\mathbf{1}_{\mathbb{N}\setminus A}(k) \geq \lambda^{-1}\Big(1 - e^{-\lambda}\Big).$$

We conclude that $\big\|U_{\mathrm{Po}(\lambda)}h\big\|_\infty \leq \lambda^{-1}\big(1 - e^{-\lambda}\big)$ for $h \in \mathcal{H}_{\mathrm{TV}}$. □

Remark 5.4. A similarly elementary but more fiddly argument can be used to show that the bound $\|f\|_\infty \leq \min\big\{1, 1.4\lambda^{-1/2}\big\}$ is also valid (see [BE83, Lemma 4]). A much more sophisticated argument in [BHJ92, Remark 10.2.4] gives the sharper bound $\|f\|_\infty \leq \min\Big\{1, \sqrt{2/(e\lambda)}\Big\}$.

We will see some examples of Poisson approximation using Stein's method in the coming sections.

6 Size-Bias Coupling

In both the Normal and Poisson cases, estimation of the necessary expectation $\mathbb{E}_X T_0 f$ involves some control over expectations of the form $\mathbb{E}[Xf(X)]$. In Chapter 7, we were able to achieve this with an ad-hoc argument exploiting independence in our random variable of interest. However, in order to apply Stein's method in less ideal circumstances, we will need some more machinery. There are a number of different techiques that can be used in different circumstances. In the case of positive random variables such as those found in combinatorial applications, one particularly useful technique is *size-bias coupling*, which we will discuss in this section.

Definition 6.1. The *size-bias distribution* $\mathcal{L}_X^{(s)}$ with respect to a nonnegative integrable random variable X is given by $\mathcal{L}_X^{(s)}(A) = \mathbb{E}[X\,\mathbf{1}_{X\in A}]/\mathbb{E}X$. In this section, $X^{(s)}$ will always denote a random variable with the distribution $\mathcal{L}_X^{(s)}$.

Remark 6.2. The reason this is called size-biasing is because the probability of each value of X is biased proportionally by that value. To be precise, in the discrete the definition of $\mathcal{L}_X^{(s)}$ can be stated as

$$\mathcal{L}_X^{(s)}(k) = \frac{k}{\mathbb{E}X}\mathcal{L}_X(k). \tag{6.1}$$

for each k. This often has a combinatorial interpretation. Intuitively speaking, if $\mathcal{L}_X$ is the number of balls in a random bin, then $\mathcal{L}_X^{(s)}$ is the distribution of the number of balls in the bin of a random ball.

We can immediately see from Remark 6.2 that in the discrete case

$$\mathbb{E}f\left(X^{(s)}\right) = \mathbb{E}[Xf(X)]/\mathbb{E}X \tag{6.2}$$

for all $\mathcal{L}_X^{(s)}$-integrable functions f. This is in fact true in general, which can be straightforwardly proved with the measure-theoretic definition of the integral.

Size-biasing therefore gives us a way to work with expectations of the form $\mathbb{E}[Xf(X)]$. In order to use this for Stein's method, we need to couple X with $X^{(s)}$ in an appropriate way. There is

a general construction that can be frequently used for both normal and Poisson approximation in combinatorial applications.

Proposition 6.3. *Let $(Q_i)_{i=1}^n$ be a sequence of zero-one random variables, and let $X = \sum_{i=1}^n Q_i$. Suppose $\left(X^{(i)}\right)_{i=1}^n, (Q_i)_{i=1}^n$ and X are coupled in such a way that $X^{(i)} \in \mathcal{L}(X|Q_i = 1)$ for each i. Let I be a random variable independent of all the random variables defined so far, with distribution satisfying $\mathcal{L}_I(i) = \mathbb{P}(Q_i = 1)/\mathbb{E}X$. Then $X^{(I)} \in \mathcal{L}_X^{(s)}$.*

Proof. We have

$$\begin{aligned}\mathbb{P}\Big(X^{(I)} = k\Big) &= \sum_{i=1}^n \frac{\mathbb{P}(Q_i = 1)}{\mathbb{E}X}\mathbb{P}(X = k|Q_i = 1) \\ &= \frac{1}{\mathbb{E}X}\sum_{i=1}^n \mathbb{P}(X = k \text{ and } Q_i = 1).\end{aligned}$$

Now, when $X = k$, exactly k of the events $\{Q_i = 1\}$ will hold, so

$$\sum_{i=1}^n \mathbb{P}(X = k \text{ and } Q_i = 1) = \mathcal{L}_X(k).$$

Then, the discrete definition (6.1) for the size-bias distribution proves $X^{(I)} \in \mathcal{L}_X^{(s)}$. □

Example 6.4 (adapted from [Ros11, Example 4.21])**.** We say a vertex in a graph is *isolated* if it has no neighbours. Fix n and p, and let $G \in \mathcal{G}_{n,p}$ (see Section 2.7). Let

$$Q_i = \mathbf{1}\{i \text{ is an isolated vertex in } G\},$$

so that $X = \sum_{i=1}^n Q_i$ is the number of isolated vertices in G. Now, the event $Q_i = 1$ means that i has no incident edges. By the independence in the definition of $\mathcal{G}_{n,p}$, the distribution of $\mathcal{L}(X|Q_i = 1)$ can be realized as the number of isolated vertices in a graph obtained by deleting from G all edges incident to i. By symmetry, each $\mathbb{P}(Q_i = 1)$ is equal, so we can couple $X^{(s)}$ with X by letting $X^{(s)}$ be the number of isolated vertices in the graph obtained by picking a vertex in G at random and deleting all its adjacent edges.

Now, the distribution of X depends on the relationship between n and p. First, note that each Q_i is "almost" independent. Next, the probability that a vertex is isolated is $(1-p)^{n-1}$

because the $n-1$ potential edges incident to that vertex must independently not be present. So if p is not vanishingly small compared to n then isolated points are a "rare event" and we might expect X to have an approximate Poisson distribution. If p vanishes at just the right rate (compared to n), then we might expect the central limit theorem to approximately hold, so that X has an approximate normal distribution. We will therefore return to this example to illustrate the application of size-bias coupling for both normal and Poisson approximation.

In both cases, we will need to compute $\mathbb{E}X$ and $\operatorname{Var} X$, so we do this now. Using Remark 2.17, we can immediately deduce that $\mathbb{E}X = n(1-p)^{n-1}$. For a pair of distinct vertices, the probability that both are isolated is $(1-p)^{2n-3}$, so

$$\mathbb{E}[X(X-1)/2] = \binom{n}{2}(1-p)^{2n-3}$$

and

$$\begin{aligned}\operatorname{Var} X &= n(n-1)(1-p)^{2n-3} + n(1-p)^{n-1} - n^2(1-p)^{2n-2} \\ &= n(1-p)^{n-1}\Big(1 + (np-1)(1-p)^{n-2}\Big).\end{aligned}$$

6.1 Normal approximation

For a nonnegative random variable X, let $\mu = \mathbb{E}X$ and $\sigma^2 = \operatorname{Var} X$. Suppose we believe $W = (X-\mu)/\sigma$ is approximately distributed as $\mathcal{N}(0,1)$. In order to prove this, our objective is to compare $\mathbb{E}[f(W)]$ to $\mathbb{E}[f'(W)]$, as in Section 4.1.

Now, for any function $f : \mathbb{R} \to \mathbb{R}$, by considering (6.2) with the function $x \mapsto f((X-\mu)/\sigma)$, we have

$$\mathbb{E}[Wf(W)] = \mathbb{E}\left[\frac{X-\mu}{\sigma} f\left(\frac{X-\mu}{\sigma}\right)\right] = \frac{\mu}{\sigma}\mathbb{E}\left[f\left(\frac{X^{(s)}-\mu}{\sigma}\right) - f\left(\frac{X-\mu}{\sigma}\right)\right].$$

A Taylor expansion (assuming f is twice-differentiable) then gives

$$\mathbb{E}[Wf(W)] = \frac{\mu}{\sigma}\mathbb{E}\left[\frac{X^{(s)}-X}{\sigma} f'\left(\frac{X-\mu}{\sigma}\right) + \frac{1}{2}\left(\frac{X^{(s)}-X}{\sigma}\right)^2 f''(E)\right],$$

for some (random) E. So, for $T_0 = T_{\mathcal{N}}$ as defined in Theorem 4.1, we have

$$|\mathbb{E}_W T_0 f| \leq \frac{3\mu}{\sigma^2}\sqrt{\frac{\pi}{2}}\mathbb{E}\left|\frac{\sigma^2}{\mu} - \mathbb{E}^X\left[X^{(s)} - X\right]\right| + \frac{3\mu}{\sigma^3}\mathbb{E}\Big(X^{(s)} - X\Big)^2. \tag{6.3}$$

for f satisfying the bounds in Theorem 4.3. (We have used the tower law of expectation, see Proposition 2.23). Now, by considering the function $x \mapsto x$ in (6.2), observe that

$$\mathbb{E}X^{(s)} = \frac{\mathbb{E}X^2}{\mathbb{E}X} = \frac{\sigma^2 + \mu^2}{\mu},$$

so $\mathbb{E}\big[\mathbb{E}^X\big[X^{(s)} - X\big]\big] = \mathbb{E}\big[X^{(s)} - X\big] = \sigma^2/\mu$. Hölder's inequality (see [Rud66, Theorem 3.6 (1)]) for the L^2 norm then gives

$$\mathbb{E}\left|\frac{\sigma^2}{\mu} - \mathbb{E}^X\left[X^{(s)} - X\right]\right| \leq \sqrt{\mathrm{Var}\big(\mathbb{E}^X\big[X^{(s)} - X\big]\big)}. \tag{6.4}$$

It follows that $d_{\mathrm{W}}(X, \mathcal{N}(0,1))$ will be small when $X^{(s)}$ is coupled in such a way that it is a slight perturbation of X. We will show how $d_{\mathrm{W}}(X, \mathcal{N}(0,1))$ can be bounded in practice with the continuation of Example 6.4.

Example 6.5. As in Example 6.4, let X be the number of isolated vertices in a random $G \in \mathcal{G}_{n,p}$, let I be a random vertex, and let $X^{(s)}$ be the number of isolated vertices with the edges incident to a random vertex deleted. For each vertex i in G, let D_i be the number of vertices adjacent to i which have degree 1, and as before let Q_i be the indicator random variable for the event that i is isolated. Then, we have

$$X^{(s)} - X = D_I + Q_I.$$

First, note that

$$\mathbb{E}\Big(X^{(s)} - X\Big)^2 = \sum_{i=1}^{n}\frac{1}{n}\mathbb{E}(D_i + Q_i)^2 \leq \sum_{i=1}^{n}\frac{1}{n}\mathbb{E}(D_i + 1)^2 = \mathbb{E}D_i^2 + 2\mathbb{E}D_i + 1. \tag{6.5}$$

Next, by contractivity and the tower law (see Proposition 2.23), we have

$$\mathrm{Var}\Big(\mathbb{E}^X\Big[X^{(s)} - X\Big]\Big) = \mathrm{Var}\Big(\mathbb{E}^X\Big[\mathbb{E}^G\Big[X^{(s)} - X\Big]\Big]\Big) \leq \mathrm{Var}\Big(\mathbb{E}^G\Big[X^{(s)} - X\Big]\Big). \tag{6.6}$$

The interpretation of $\mathbb{E}^G\big[X^{(s)} - X\big]$ is that $X^{(s)} - X$ is averaged over all choices of the random vertex I, for each G. That is,

$$\mathbb{E}^G\Big[X^{(s)} - X\Big] = \sum_{i=1}^{n} \frac{1}{n}(D_i + Q_i).$$

Now, let $D = \sum_{i=1}^{n} D_i$ be the number of vertices with degree 1 in G, and let $\hat{D} = D - \mathbb{E}D$. Recall that $X = \sum_{i=1}^{n} Q_i$, and let $\hat{X} = X - \mathbb{E}X$. We have

$$\operatorname{Var}\Big(\mathbb{E}^G\Big[X^{(s)} - X\Big]\Big) = \frac{1}{n^2}\mathbb{E}\Big(\hat{D} + \hat{X}\Big)^2 = \frac{1}{n^2}\mathbb{E}\Big[2\hat{D}^2 - 2\hat{X}^2 - \Big(\hat{D} - \hat{X}\Big)^2\Big] \leq \frac{2}{n^2}(\operatorname{Var} D + \operatorname{Var} X). \tag{6.7}$$

Combining Equations (6.3), (6.4), (6.5), (6.6) and (6.7), we just need to estimate $\operatorname{Var} D$, $\mathbb{E}D_i$ and $\mathbb{E}D_i^2$ in order to obtain a bound for

$$d_{\mathrm{W}}\left(\frac{X - \mu}{\sigma}, \mathcal{N}(0,1)\right).$$

6.2 Poisson Approximation

7 The Method of Exchangeable Pairs

In [Ste86], Stein introduced a method for the construction of a characterizing operator using an object called an *exchangeable pair*. Up until this point, we have only been interested in characterizing operators for our "special" distributions $\mathcal{L}_0$ to which we compare our distribution of interest $\mathcal{L}_X$. The exchangeable pairs construction can be used for this purpose, but in this section we will be more interested in a technique for applying Stein's method that involves the construction of a characterizing operator for $\mathcal{L}_X$.

The general idea is that we define an exchangeable pair $\mathbf{X}$, and by a general construction this gives a characterizing operator $T_{\mathbf{X}}$ for our distribution of interest $\mathcal{L}_X$. We then use an operator α to connect the domains of $T_{\mathbf{X}}$ and T_0 in such a way that $T_{\mathbf{X}}\alpha$ approximates T_0. By the definition of a characterizing operator, we have $\mathbb{E}_X T_0 = \mathbb{E}_X(T_0 - T_{\mathbf{X}}\alpha)$, so we can use the fact that $T_0 - T_{\mathbf{X}}\alpha$ is small to bound $\mathbb{E}_X T_0 f$.

This method has been applied in some generality, but for simplicity, we assume throughout that the random variable of interest X has finite support Ω_X.

Example 7.1 (adapted from [Ros11, Example 4.21])**.** Throughout this section, we will refer to an example problem to illustrate the principles of Stein's method for exchangeable pairs.

We will say a *fixed point* of a permutation $\sigma \in S_n$ is an index $k \in [n]$ that satisfies $\sigma(k) = k$. Let $X : S_n \to \{0\} \cup [n]$ give the number of fixed points in each permutation from S_n. We interpret X as a random variable on the underlying uniform space $\mathcal{S}_n$.

Now, if n is large then fixed points are largely independent of each other, and each of n indices has a probability of $1/n$ to be a fixed point. So (recalling Remark 2.17), we might expect $\mathcal{L}_X$ to be "close" to Po(1). We will attempt to bound $d_{\mathrm{TV}}(X, \mathrm{Po}(1))$ to validate and quantify this intuition.

Definition 7.2. A 2-dimensional random pair $\mathbf{X} = (X_1, X_2)$ is an *exchangeable pair* if $\mathcal{L}(X_1, X_2) = \mathcal{L}(X_2, X_1)$. We will denote the support of $\mathbf{X}$ by $\Omega_{\mathbf{X}}^{(2)}$ (this is a subset of Ω_X^2).

That is, a pair $\mathbf{X}$ is exchangeable if exchanging the components of the pair does not change their joint distribution. In particular, the marginal distributions of X_1 and X_2 must be the same.

Proposition 7.3. *There is a natural equivalence between time-homogeneous reversible Markov Chains with steady-state distribution $\mathcal{L}_X$, and exchangeable pairs with margins $\mathcal{L}_X$.*

Proof. Given an exchangeable pair $\mathbf{X}$ with margins $\mathcal{L}_X$, we can define a time-homogenous Markov chain M with transition probabilities $p(x_1, x_2) = \mathbb{P}(X_2 = x_2 | X_1 = x_1)$. With $\pi(x) = \mathcal{L}_X(x)$, we then have

$$\pi(x_1)p(x_1, x_2) = \mathcal{L}_{\mathbf{X}}(x_1, x_2) = \mathcal{L}_{\mathbf{X}}(x_2, x_1) = \pi(x_2)p(x_2, x_1)$$

for any $x_1, x_2 \in \Omega_X$. So, M is reversible with steady-state distribution $\mathcal{L}_X$. Of course, given a reversible Markov Chain with steady-state distribution $\mathcal{L}_X$, we can define the distribution $\mathcal{L}_{\mathbf{X}}$ in exactly the same way.

□

Definition 7.4. We say an exchangeable pair $\mathbf{X}$ is *connected* if the corresponding Markov Chain is irreducible.

Remark 7.5. We are particularly interested in exchangeable pairs $\mathbf{X}$ with marginal distributions $\mathcal{L}_{X_1} = \mathcal{L}_{X_2} = \mathcal{L}_X$. If X is defined on an underlying combinatorial probability space $\big(\Omega, 2^\Omega, \mathbb{P}\big)$, it is often convenient to first construct an exchangeable pair $\mathbf{W} = (W_1, W_2)$ with margins $\mathbb{P}$, so that the vector $\mathbf{X_W} = (X(W_1), X(W_2))$ is an exchangeable pair with margins $\mathcal{L}_X$. If (W_1, W_2) is connected, then $\mathbf{X_W}$ is connected also.

Example 7.6. We continue Example 7.1. We will define a specific exchangeable pair $\mathbf{W} = (W_1, W_2)$ with margins $\mathcal{S}_n$ by

$$\mathbb{P}((W_1, W_2) = (\sigma_1, \sigma_2)) = \begin{cases} \big(n!\binom{n}{2}\big)^{-1} & \text{if } \sigma_1 = \sigma_2(i\,j) \text{ for some transposition } (i\,j), \\ 0 & \text{otherwise.} \end{cases}$$

The relation of differing by a transposition is symmetric, so $\mathbf{W}$ is indeed an exchangeable pair. The Markov Chain associated with $\mathbf{W}$ has a simple interpretation. Given a random permutation σ, to make a transition in the Markov Chain we just randomly choose one of the $\binom{n}{2}$ possible transpositions and compose it with σ. Because the transpositions generate S_n, the pair $\mathbf{W}$ is connected, so we can use the construction from Remark 7.5 to produce a connected exchangeable pair $\mathbf{X}$ with margins X.

The Markov Chain underlying a connected exchangeable pair can be naturally viewed as a connected one-dimensional simplicial complex. The zeroth reduced homology group $\ker \partial_0 / \operatorname{im} \partial_1$ of a connected simplicial complex has dimension zero, and this motivates the construction of a characterizing operator in a natural way. (The following theorem is self-contained and requires no knowledge of homology).

Theorem 7.7. *Suppose* $\mathbf{X}$ *is a connected exchangeable pair with margins* $\mathcal{L}_X$*. Let* $\mathcal{F}_X$ *be the set of functions* $f : \Omega_X^2 \to \mathbb{R}$ *which are antisymmetric in the sense that* $f(x_1, x_2) = -f(x_2, x_1)$*. Let* $\mathcal{V}_X$ *be the set of functions* $h : \Omega_X \to \mathbb{R}$*.*

Define $T_{\mathbf{X}} : \mathcal{F}_X \to \mathcal{V}_X$ *by* $T_{\mathbf{X}} f(x) = \sum_{x_2 \in \Omega_X} f(x, x_2) p(x, x_2) = \mathbb{E}[f(\mathbf{X}) | X_1 = x]$*, so that* $T_{\mathbf{X}} f(X)$ *is distributed as* $\mathbb{E}^{X_1} f(\mathbf{X})$*. Then* $T_{\mathbf{X}}$ *is a characterizing operator for* X*.*

Remark 7.8. This theorem appears in [Ste92, Section 4], but the proof appears to be incorrect. I have adjusted it fairly significantly.

Proof. To see that $\operatorname{im} T_{\mathbf{X}} \subseteq \ker \mathbb{E}_X$, fix $f \in \mathcal{F}_X$ and note that by the tower law of expectation (Proposition 2.23),

$$\mathbb{E}_X T_{\mathbf{X}} f = \mathbb{E}\mathbb{E}^{X_1} f(\mathbf{X}) = \mathbb{E} f(\mathbf{X}).$$

By exchangeability and antisymmetry, $\mathbb{E} f(\mathbf{X}) = \mathbb{E} f(X_2, X_1) = -\mathbb{E} f(\mathbf{X})$, so $\mathbb{E} f(X_1, X_2) = \mathbb{E}_X T_{\mathbf{X}} f = 0$. This did not require the connectedness condition. We can similarly prove that $T_{\mathbf{X}}$ is well-defined as an operator from $\mathcal{F}_X$ to $\mathcal{V}_X$: note that for $f \in \mathcal{F}_X$ we have $\mathbb{E}_X |T_{\mathbf{X}} f| \leq \mathbb{E} |f(\mathbf{X})| < \infty$ by contractivity, so $T_{\mathbf{X}} f \in \mathcal{V}_X$.

We will next prove $\ker \mathbb{E}_X \subseteq \operatorname{im} T_{\mathbf{X}}$, but first we make some definitions. For each $x \in \Omega_X$, let h_x be the function that takes the value $\pi(x)^{-1}$ on x and is zero elsewhere, so that

$$h = \sum_{x \in \Omega_X} h(x) \pi(x) h_x$$

for each $h \in \mathcal{V}_X$. For each $(x_1, x_2) \in \Omega_{\mathbf{X}}^{(2)}$, define $f_{x_1, x_2} \in \mathcal{F}_X$ as the function that takes the value $(\pi(x_1) p(x_1, x_2))^{-1}$ on (x_1, x_2), takes the value $-(\pi(x_2) p(x_2, x_1))^{-1}$ on (x_2, x_1), and takes the value zero elsewhere. Note that this function is antisymmetric by the reversibility of the Markov Chain of $\mathbf{X}$. We have $T_{\mathbf{X}} f_{x_1, x_2} = h_{x_2} - h_{x_1}$.

Let $h \in \ker \mathbb{E}_X$, and fix an arbitrary $x^* \in \Omega_X$. By the connectedness assumption, for each $x \in \Omega_X$ there is a sequence

$$x = x^{(0)}, x^{(1)}, \ldots, x^{(k-1)}, x^{(k)} = x^*$$

with $\left(x^{(i-1)}, x^{(i)}\right) \in \Omega_{\mathbf{X}}^{(2)}$ for $i = 1, \ldots k$. Define

$$f_x^* = \sum_{i=1}^{k} f_{x^{(i-1)}, x^{(i)}},$$

so that $h_{x^*} - h_x = T_{\mathbf{X}} f_x^*$ by linearity. It follows that

$$h = \sum_{x \in \Omega_X} h(x)\pi(x)(h_{x^*} - T_{\mathbf{X}} f_x^*) = \sum_{x \in \Omega_X} (\mathbb{E}_X h) h_{x^*} - T_{\mathbf{X}} \sum_{x \in \Omega_X} h(x)\pi(x) f_x^* \in \operatorname{im} T_{\mathbf{X}}.$$

Once we have defined an exchangeable pair and used it to construct $T_{\mathbf{X}}$, the final step is to choose an operator $\alpha : \mathcal{F}_0 \to \mathcal{F}_X$ in such a way that T_0 can be easily compared with $T_{\mathbf{X}}\alpha$. $\square$

Comment 7.1. It's not clear if characterizing operators are in any sense unique so that for two characterizing operators T_1, T_2 there is *always* a connection α that makes $T_1 = T_2\alpha$.

Example 7.9. For the Poisson case in Theorem 5.1, we need to compare $\lambda f(X+1)$ with $Xf(X)$. It is often fruitful to define α by

$$\alpha f(x_1, x_2) = c f(x_2)\, \mathbf{1}\{x_2 = x_1 + 1\} - c f(x_1)\, \mathbf{1}\{x_1 = x_2 + 1\}$$

for some $c \in \mathbb{R}$. We will then have

$$(T_{\mathbf{X}}\alpha f)(x) = c f(x+1)\mathbb{P}(X_2 = X_1 + 1 | X_1 = x) - c f(x)\mathbb{P}(X_1 = X_2 + 1 | X_1 = x)$$

so $\mathbb{E}_X T_0 f = \mathbb{E}_X(T_0 f - T_{\mathbf{X}}\alpha f)$ equals

$$\mathbb{E}[f(X_1+1)(\lambda - c\mathbb{P}(X_2 = X_1 + 1 | X_1)) - f(X_1)(X_1 - c\mathbb{P}(X_1 = X_2 + 1 | X_1))].$$

Using the triangle inequality, we have

$$|\mathbb{E}_X T_0 f| \le \min\Big(2, 3\lambda^{-1/2}\Big)(\mathbb{E}|\lambda - c\mathbb{P}(X_2 = X_1 + 1 | X_1)| + \mathbb{E}|X_1 - c\mathbb{P}(X_1 = X_2 + 1 | X_1)|) \tag{7.1}$$

for all f satisfying the bound in Theorem 5.3.

This approximation is effective when

$$\mathbb{P}(X_2 = X_1 + 1 | X_1) \approx \lambda/c, \quad \mathbb{P}(X_2 = X_1 - 1 | X_1) \approx X_1/c. \tag{7.2}$$

The interpretation of these approximate equalities is that the Markov Chain associated with $\mathbf{X}$ is approximately an immigration-death process. This is likely to happen when $X(\omega)$ is in some sense a statistic of the amount of local structure over the object ω, and $\mathbf{X}$ is defined by a Markov Chain on Ω (as in Remark 7.5) that (uniformly) randomly disturbs local structure. The conclusion to Example 7.1, presented in Example 7.10 below, should make this clear.

Example 7.10. We continue with Example 7.1, recalling the exchangeable pairs $\mathbf{W}$ and $\mathbf{X}$ from Example 7.6. The interpretation of

$$\mathbb{P}(X_2 = X_1 - 1|X_1)$$

is the probability of a transposition destroying exactly one out of the existing X_1 fixed points. In order to destroy exactly one fixed point, we have to choose a fixed point to destroy, and swap it with a non-fixed-point. There are $X_1(n - X_1)$ out of $\binom{n}{2}$ transpositions that do this, so

$$\mathbb{P}(X_2 = X_1 - 1|X_1) = \frac{X_1(n - X_1)}{\binom{n}{2}}.$$

Next, we will find a formula for $\mathbb{P}(X(W_2) = X(W_1) + 1|W_1)$, noting that

$$\mathbb{P}(X_2 = X_1 + 1|X_1) = \mathbb{E}[\mathbb{P}(X(W_2) = X(W_1) + 1|W_1)|X_1].$$

In order to create exactly one new fixed point, we have to choose an index k that is not fixed in W_1 (there are $n - X_1$ such) and compose W_1 with the transposition $(k\ \sigma(k))$. This creates exactly one fixed point unless $\sigma^{-1}(k) = k$, in which case it creates two. We have counted this second case twice for every transposition in the cycle decomposition of W_1. Let Y be the number of transpositions in the cycle decomposition of W_1. We have

$$\mathbb{P}(X_2 = X_1 + 1|X_1) = \frac{n - X_1 - 2\mathbb{E}[Y|X_1]}{\binom{n}{2}}.$$

In order to satisfy (7.2) as closely as possible, we choose $c = \binom{n}{2}/n$. Recalling that $\mathbb{E}X_1 = 1$,

we then have

$$\begin{aligned}\mathbb{E}|1 - c\mathbb{P}(X_2 = X_1 + 1|X_1)| &= \mathbb{E}\left[1 - \frac{n - X_1 - 2\mathbb{E}[Y|X_1]}{n}\right] \\ &= 1/n + 2\mathbb{E}Y/n, \qquad (7.3) \\ \mathbb{E}|X_1 - c\mathbb{P}(X_2 = X_1 - 1|X_1)| &= \mathbb{E}\left[X_1 - \frac{X_1(n - X_1)}{n}\right] \\ &= \mathbb{E}X_1^2/n. \qquad (7.4)\end{aligned}$$

Now, the probability that a transposition $(i\,j)$ is in the cycle decomposition of W_1 is $(n(n-1))^{-1}$ because i must map to j out of the n possible options in $[n]$, then j must map to i out of the $n-1$ possible options in $[n]\backslash\{j\}$. There are $\binom{n}{2} = n(n-1)/2$ possible transpositions so, by Remark 2.17, it follows that $\mathbb{E}Y = 1/2$.

Now, $\mathbb{E}[X_1(X_1-1)/2]$ is the expected number of unordered pairs of distinct fixed points in a permutation $\sigma \in \mathcal{S}_n$. For any unordered pair of distinct indices $\{i, j\}$, the probability that both are fixed is $(n(n-1))^{-1}$ because i must map to i out of the n possible options in $[n]$, then j must map to j out of the $n-1$ possible options in $[n]\backslash\{i\}$. The total number of unordered pairs is $\binom{n}{2} = n(n-1)/2$, so again applying Remark 2.17, we have $\mathbb{E}[X_1(X_1-1)/2] = 1/2$ and $\mathbb{E}\big[X_1^2\big] = 2$.

Combining (3.4), (7.1), (7.3) and (7.4), we can conclude that $d_{\mathrm{TV}}(X, \mathrm{Po}(1)) \leq 8/n$.

We stress that this example was included only for illustrative purposes. Since d_{TV} has a relatively direct probabilistic interpretation, and since permutations are very simple, well-understood combinatorial objects, it is actually possible to analyze $d_{\mathrm{TV}}(X, \mathrm{Po}(1))$ with elementary (but complicated) combinatorics. One can obtain super-exponential bounds which are much better than the bound we obtained by Stein's method (see [ABT03, Section 1.1]). However, such elementary analysis is sometimes not possible, as we will now see.

7.1 Switchings and Short Cycles in Random Regular Graphs

A standard way to enumerate combinatorial objects is by way of a recurrence. That is, we try to show that every object can be built up from independent substructures, so the number of

objects of a given size is some function of the number of objects of smaller sizes. The most difficult combinatorial structures to enumerate and study, then, are those that are constrained in some way so that they exhibit self-dependence, meaning that they cannot be built up from substructures. For example, regular graphs are graphs where each vertex has the same degree. There is no simple, uniform way to construct a regular graph from regular graphs of smaller degree or with fewer vertices.

Combinatorial *switchings* are a powerful tool for the analysis of these kinds of combinatorial structures. In particular, they have been applied with great success to the enumeration of regular graphs and Latin squares.

A particularly interesting application of Stein's method for exchangeable pairs is to piggyback onto results proved using switchings. This idea was very recently introduced by Johnson [Joh11]. In this subsection, I'll describe how Johnson improved a certain Poisson Limit Theorem concerning short cycles in random regular graphs, in a quite natural, and generally applicable way. First, we give some background on short cycles in random regular graphs, and switchings.

7.1.1 Short Cycles in Random Regular Graphs

Let $X_{\ell,n}^{(d)}$ be the number of cycles of length ℓ in a random d-regular graph on n vertices. For all asymptotics we will assume $n \to \infty$ in such a way that nd is always even (otherwise there are no d-regular graphs on n vertices). The following is a critical theorem describing the structure of random regular graphs.

Theorem 7.11. *Fix $\ell \in \mathbb{N}$ and $d \in \mathbb{N}$ with $d \geq 3$. Then*

$$\mathcal{L}\left(X_{\ell,n}^{(d)}\right) \to \mathrm{Po}\left(\frac{(d-1)^{\ell}}{2\ell}\right)$$

as $n \to \infty$.

This theorem is most easily proved with the method of moments; see [Wor99, Theorem 2.5].

Remark 7.12. The full theorem is that for any $r \in \mathbb{N}$, the vector of short cycle counts $\left(X_{1,n}^{(d)}, \ldots X_{r,n}^{(d)}\right)$ is asymptotically distributed as an *independent* vector of Poisson random variables. That is, not only do we understand the distributions of the short cycle counts, we also know that these cycle counts do not interact with each other, asymptotically speaking. Since we have not discussed the weak topology on random vectors, all the results and discussion in this subsection are simplified to the single-variable case.

Remark 7.13. One reason Theorem 7.11 is interesting is because it tells us the number of "short cycles" in a random regular graph does not grow (to infinity) with the size of the graph. That is, random regular graphs are likely to "locally" "look like forests".

Theorem 7.11 can be extended to the case where d is allowed to grow modestly with n. In [MWW04], a natural variant of Theorem 7.11 was proved for $(d-1)^{2\ell-1} = o(n)$, using the idea of *switchings*. The condition $(d-1)^{2\ell-1} = o(n)$ appeared to be a natural threshold; it was conjectured that short cycle counts are no longer asymptotically Poisson if d grows any faster.

It was recently proved that in fact cycle counts remain asymptotically Poisson past this boundary:

Theorem 7.14 ([Joh11])**.** *Let ℓ and d depend on n, in such a way that $d \geq 3$ and*

$$\sqrt{\ell}(d-1)^{3\ell/2-1} = o(n).$$

Let $n \to \infty$ in such a way that nd is always even. Then,

$$d_{\mathrm{TV}}\left(X_{\ell,n}^{(d)}, \mathrm{Po}\left(\frac{(d-1)^{\ell}}{2\ell}\right)\right) \to 0.$$

as $n \to \infty$.

The proof of Theorem 7.14 combines switchings with Stein's method.

7.1.2 Switchings

Formally speaking, a switching is a binary relation $\rightsquigarrow$ on a set of combinatorial objects Ω (in our case, Ω is the set of d-regular graphs on n vertices). Usually, a switching is understood as an "action" that changes one combinatorial object to another. The switching used in [MWW04] is defined as follows:

Definition 7.15. Let $C = v_1 \dots v_\ell v_1$ be a cycle of length ℓ in a d-regular graph G. For each $i \in \mathbb{Z}/\ell\mathbb{Z}$, suppose $e_i = u_i w_{i+1} \in E(G)$ satisfies $u_i v_i \notin E(G)$ and $v_i w_i \notin E(G)$. (Each of the 3ℓ vertices in v_i, u_i, w_i must be distinct). Let G' be the d-regular graph obtained from G by deleting each e_i and all the edges in C and adding each of the edges $u_i v_i$ and $v_i w_i$. Suppose that no cycles of length less than or equal to ℓ other than C are created or destroyed by this process. Then, we say $G \rightsquigarrow G'$. Figure 2 below illustrates the idea.

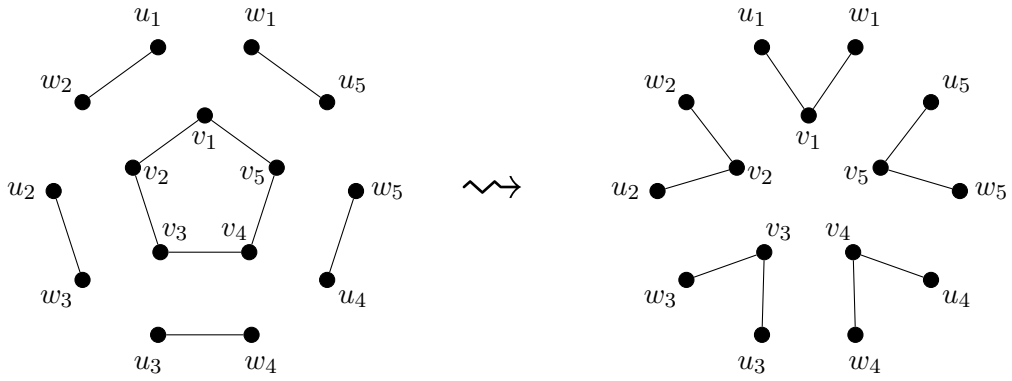

Figure 2: A switching with $\ell = 5$ (unaffected edges and vertices are not pictured).

The typical application of a switching is to estimate the relative sizes of some subsets of Ω, by analysing the "flow" of switchings between the subsets. The switching is generally designed in such a way that this "flow" is maximized in a certain "direction". See [HM10] for a discussion of switchings in generality.

For our case, define $S(k)$ to be the set of d-regular graphs with exactly k cycles of length ℓ. The switching in Definition 7.15 has been chosen so that it attempts to remove exactly one short cycle in a graph, so that for each k the "flow" between $S(k)$ and $S(k-1)$ is large.

In [MWW04, Section 3], an estimate was obtained for the average number of ways $F(k)$ to "switch into $S(k-1)$" from $S(k)$ (that is, the average number of $G' \in S(k-1)$ satisfying $G \rightsquigarrow G'$, where G ranges over $S(k)$). Similarly, an estimate was obtained for the average number of ways $B(k-1)$ to "switch from $S(k)$" into $S(k-1)$ (that is, the average number of $G \in S(k)$ satisfying $G \rightsquigarrow G'$, where G' ranges over $S(k-1)$). Then, $S(k)F(k) = S(k-1)B(k-1)$ is the number of pairs $(G, G') \in S(k) \times S(k-1)$ with $G \rightsquigarrow G'$.

The estimates on $F(k)$ and $B(k)$ can be combined for an estimate of

$$\frac{\mathbb{P}\Big(X_{\ell,n}^{(d)} = k\Big)}{\mathbb{P}\Big(X_{\ell,n}^{(d)} = 0\Big)} = \frac{S(k)}{S(0)} = \frac{B(0)\dots B(k-1)}{F(1)\dots F(k)}, \tag{7.5}$$

which can then be used to estimate the distribution of $X_{\ell,n}^{(d)}$.

7.1.3 Stein's Method and Switchings

Obvious parallels can be drawn between the application of switchings in [MWW04] and the application of Stein's method in Example 7.9. $F(k)$ can be interpreted as the "probability" that the ("forwards") switching "destroys a cycle", and $B(k)$ can be interpreted as the "probability" that the "backwards" switching "creates a cycle", both given that there are k existing cycles. These are analogous to the probabilities $\mathbb{P}(X_2 = X_1 - 1|X_1)$ and $\mathbb{P}(X_2 = X_1 + 1|X_1)$ that were estimated in Example 7.9.

The idea, then, is to use a switching to define a suitable exchangeable pair, and to adapt the switching estimates for use with Stein's method. There are two main advantages of this approach over a bare-hands switching argument. First, a regular switching argument requires individual estimates of each $F(k)$ and $B(k)$, but Stein's method allows us to estimate the relevant probabilities only in expectation (effectively, we are only estimating $\mathbb{E}[\lambda - cB(X)]$ and $\mathbb{E}[X - cF(X)]$, where λ is the parameter of the approximating Poisson distribution). Second, our final estimate is likely to be sharper when using Stein's method, because we can take advantage of the sophisticated estimates developed for Stein's method rather than using

elementary ad-hoc estimates. In particular, the improvement in Theorem 7.14 over the natural bound in [MWW04] is due to the term $\lambda^{-1/2}$ in (7.1).

The idea is that a switching $\rightsquigarrow$ defines a graph $\mathfrak{G}$ with an edge for every pair (G, G') with $G \rightsquigarrow G'$, whose edges we can weight if desired. Loops are then added as necessary to so that each vertex has the same weight Δ. By Proposition 2.27, $\mathfrak{G}$ induces a reversible Markov Chain with uniform stationary distribution (in our case $\mathcal{G}_{n,d}$), which corresponds to an exchangeable pair $\mathbf{G} = (G_1, G_2)$ with margins $\mathcal{G}_{n,d}$. With Remark 7.5, this provides us an exchangeable pair $\mathbf{X} = (X_1, X_2)$ with margins $\mathcal{L}_X$, which we use to apply Stein's method.

We give a very simplified sketch of the proof of Theorem 7.14. We use the switching in Definition 7.15. Assign all edges in $\mathfrak{G}$ a weight of 1, and let $X = X_{\ell,n}^{(d)}$. Let $F_C(G)$ be the number of ways to switch out of a graph G by destroying an ℓ-cycle $C \subseteq K_n$. Now, there are $(n)_\ell/(2\ell)$ cycles in the complete graph K_n on n vertices, so

$$\begin{aligned}\mathbb{E}|X_1 - c\mathbb{P}(X_2 = X_1 + 1|X_1)| &\leq \mathbb{E}\left|\sum_{C\in K_n} \mathbf{1}_{C\subseteq G_1} - \frac{c}{\Delta}\sum_{C\in K_n} F_C(G_1)\right| \\ &\leq \frac{(n)_\ell}{2\ell}\mathbb{E}\left|\mathbf{1}_{C\subseteq G_1} - \frac{c}{\Delta}F_C(G_1)\right|.\end{aligned}$$

Similarly, let $B_C(G)$ be the number of ways to switch into G by destroying C. We have

$$\mathbb{E}\left|\frac{(d-1)^\ell}{2\ell} - c\mathbb{P}(X_2 = X_1 + 1|X_1)\right| \leq \frac{(n)_\ell}{2\ell}\,\mathbb{E}\left|\frac{(d-1)^\ell}{(n)_\ell} - \frac{c}{\Delta}B_C(G_1)\right|.$$

Now, if $C \subseteq G$, then

$$F_C(G) \approx (n)_\ell\, d^\ell/2^\ell. \tag{7.6}$$

To see this, note that a forwards switching is determined by the ℓ vertices $\{w_i\}_i$ and a neighbour u_i for each. For large n, there are about $(n)_\ell$ ways to choose the $\{w_i\}_i$, and each w_i has d neighbors (for large n these are unlikely to interfere with each other). Because each u_i and w_i can be interchanged, resulting in the same switching, we then need to divide by 2^ℓ. Also, we have

$$B_C(G) \approx \binom{d}{2}^\ell, \tag{7.7}$$

because a switching that destroyed C to create G can be identified with a path $u_iv_iw_i$ for each

$v_i \in C$, and there are about $\binom{d}{2}^\ell$ ways to choose these paths. With these estimates in mind, we choose $c = \Delta 2^\ell / \big((n)_\ell d^\ell\big)$, so that both of

$$\mathbb{E}\Big|\mathbf{1}_{C\subseteq G_1} - \frac{c}{\Delta} F_C(G_1)\Big|, \quad \mathbb{E}\left|\frac{(d-1)^\ell}{(n)_\ell} - \frac{c}{\Delta} B_C(G_1)\right| \tag{7.8}$$

are small. This will give us a bound on

$$\mathbb{E}|X_1 - c\mathbb{P}(X_2 = X_1 + 1 | X_1)| + \mathbb{E}\left|\frac{(d-1)^\ell}{2\ell} - c\mathbb{P}(X_2 = X_1 + 1 | X_1)\right|,$$

as is required to prove that

$$d_{\mathrm{TV}}\left(X, \mathrm{Po}\left(\frac{(d-1)^\ell}{(n)_\ell}\right)\right) \approx 0.$$

The arguments needed to actually obtain bounds on the expectations in (7.8) can be mostly adapted from corresponding arguments in the bare-hands application of switchings in [MWW04], which we will now sketch.

By the argument used to establish (7.6), we know that a switching that destroys a cycle C is determined by a choice of a set of edges $\{w_i u_i\}_{i=1}^\ell$. To estimate $\mathbb{E}[F_C(G)]$, we need a lower bound (in expectation) of the number of such choices that are "legal" as per Definition 7.15 (that is, the relevant vertices are distinct and no extraoneous cycles are created or destroyed). A choice will certainly be legal if we require that each of the edges $w_i u_i$ are a certain minimum distance (say ℓ) away from each other and of the cycle C, and we also require that none of the edges $w_i u_i$ are part of a "short cycle" (a cycle with length less than or equal to ℓ). The number of vertices with distance less than or equal to ℓ from a particular vertex is at most d^ℓ, which is a vanishingly small proportion of the vertices of a large graph. Due to this and the fact that short cycles are "rare" in expectation (Remark 7.13), it can be shown combinatorially that

$$\mathbb{E}\Big|\mathbf{1}_{C\subseteq G_1} - \frac{c}{\Delta} F_C(G_1)\Big| = O\left(\frac{\ell(d-1)^{2\ell-1}}{n^{\ell+1}}\right).$$

We similarly need to investigate the sharpness of (7.7) by bounding the number of "legal"

choices of sets of paths $\{u_i v_i w_i\}_{i=1}^{\ell}$. For this, we restrict our attention to such choices where no edge $v_i u_i$ or $v_i w_i$ is contained in a short cycle, and the distance between $u_i v_i w_i$ and $u_j v_j w_j$ is exactly the distance between v_i and v_j. We can then obtain a similar bound

$$\mathbb{E}\left|\frac{(d-1)^{\ell}}{(n)_{\ell}} - \frac{c}{\Delta} B_C(G_1)\right| = O\left(\frac{\ell(d-1)^{2\ell-1}}{n^{\ell+1}}\right).$$

Theorem 7.14 can be shown to follow.

8 Further Reading

There are many aspects of Stein's method that we were unable to cover in this thesis. We have discussed approximation only with the Poisson and normal distributions, but Stein's method has been applied in much more generality. The method has been applied for other "special" distributions such as the exponential, geometric and semicircle distributions ([Ros11, FG14]). More interestingly, Stein's method has been used to approximate with application-specific distributions (see, for example, [BČX07]). This is done using general characterizing operator constructions such as the exchangeable pairs approach (Theorem 7.7), and the "generator approach" ([BC05]).

Also, while we were able to discuss the method of size-bias coupling (Chapter 6), there are several other distributional transformations that can be used to control the expectations produced by the Stein transformation. The most notable of these is zero-biasing (see, for example, [CGS10, Section 2.3.3]).

We have also only discussed Stein's method as a tool for bounding the distance between distributions, but the method can be applied to prove other types of results, including, for example, concentration inequalities (see, for example, [Ros11, Section 7]). The method has even been used for totally non-probabilistic results, such as the asymptotic enumeration of Latin rectangles ([Ste86, Section XI]).

Finally, it is worth remarking that Stein's method is currently an active area of research, particularly in the fields of stochastic processes and random matrix theory. In fact, several

interesting papers appeared on the arXiv too recently for me to be able to discuss them in the body of this thesis. The reader may be interested in a very recent survey paper [Cha14] discussing the history of Stein's method and some current open problems.

References

[ABT03] R. Arratia, A. D. Barbour, and S. Tavaré, *Logarithmic Combinatorial Structures: a Probabilistic Approach*, EMS Monographs in Mathematics, European Mathematical Society, 2003.

[BC05] A. D. Barbour and L. H. Y. Chen, *An Introduction to Stein's Method*, Lecture notes series, Institute for Mathematical Sciences, National University of Singapore, vol. 4, Singapore University Press and World Scientific, 2005.

[BČX07] A. D. Barbour, V. Čekanavičius, and A. Xia, *On Stein's method and perturbations*, Latin American Journal of Probability and Mathematical Statistics **3** (2007), 31–53.

[BE83] A. Barbour and G. Eagleson, *Poisson approximation for some statistics based on exchangeable trials*, Advances in Applied Probability **15** (1983), no. 3, 585–600.

[Ber41] A. C. Berry, *The accuracy of the Gaussian approximation to the sum of independent variates*, Transactions of the American Mathematical Society **49** (1941), no. 1, 122–136.

[BHJ92] A. D. Barbour, L. Holst, and S. Janson, *Poisson Approximation*, Clarendon Press Oxford, 1992.

[CGS10] L. H. Y. Chen, L. Goldstein, and Q.-M. Shao, *Normal Approximation by Stein's Method*, Probability and its Applications, Springer, 2010.

[Cha14] S. Chatterjee, *A short survey of Stein's method*, arXiv preprint arXiv:1404.1392 (2014).

[Die00] R. Diestel, *Graph Theory*, Graduate Texts in Mathematics, Springer-Verlag, 2000.

[Ess42] C.-G. Esseen, *On the Liapounoff Limit of Error in the Theory of Probability*, Arkiv för Matematik, Astronomi och Fysik **A28** (1942), no. 9, 1–19.

[FG14] J. Fulman and L. Goldstein, *Stein's method, semicircle distribution, and reduced decompositions of the longest element in the symmetric group*, arXiv preprint arXiv:1405.1088 (2014).

[HM10] M. Hasheminezhad and B. D. McKay, *Combinatorial estimates by the switching method*, Contemporary Mathematics **531** (2010), 209–221.

[Joh11] T. Johnson, *Exchangeable pairs, switchings, and random regular graphs*, arXiv preprint arXiv:1112.0704 (2011).

[Kal02] O. Kallenberg, *Foundations of Modern Probability*, Springer, 2002.

[MWW04] B. D. McKay, N. C. Wormald, and B. Wysocka, *Short cycles in random regular graphs*, Electronic Journal of Combinatorics **11** (2004), no. R66.

[Ros11] N. Ross, *Fundamentals of Stein's method*, Probability Surveys **8** (2011), 210–293.

[Rud66] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, 1966.

[Rud73] ———, *Functional Analysis*, McGraw-Hill, 1973.

[Ste86] C. Stein, *Approximate Computation of Expectations*, IMS Lecture Notes – Monograph Series, no. 7, IMS, 1986.

[Ste92] ———, *A way of using auxiliary randomization*, Probability Theory (Singapore, 1989) (1992), 159–180.

[Ste12] S. Sternberg, *Berry-Esseen via Stein*, Problem set 3 for Math 212a, Harvard, http://isites.harvard.edu/fs/docs/icb.topic1129937.files/ex2121203.pdf, 2012.

[Vil03] C. Villani, *Topics in Optimal Transportation*, Graduate Studies in Mathematics, no. 58, American Mathematical Society, 2003.

[Wor99] N. C. Wormald, *Models of random regular graphs*, Surveys in Combinatorics (D. A. Preece and J. D. Lamb, eds.), London Mathematical Society Lecture Note Series, Cambridge University Press, 1999, pp. 239–298.